数据思维：
让大脑更清醒



▲ 套装4册

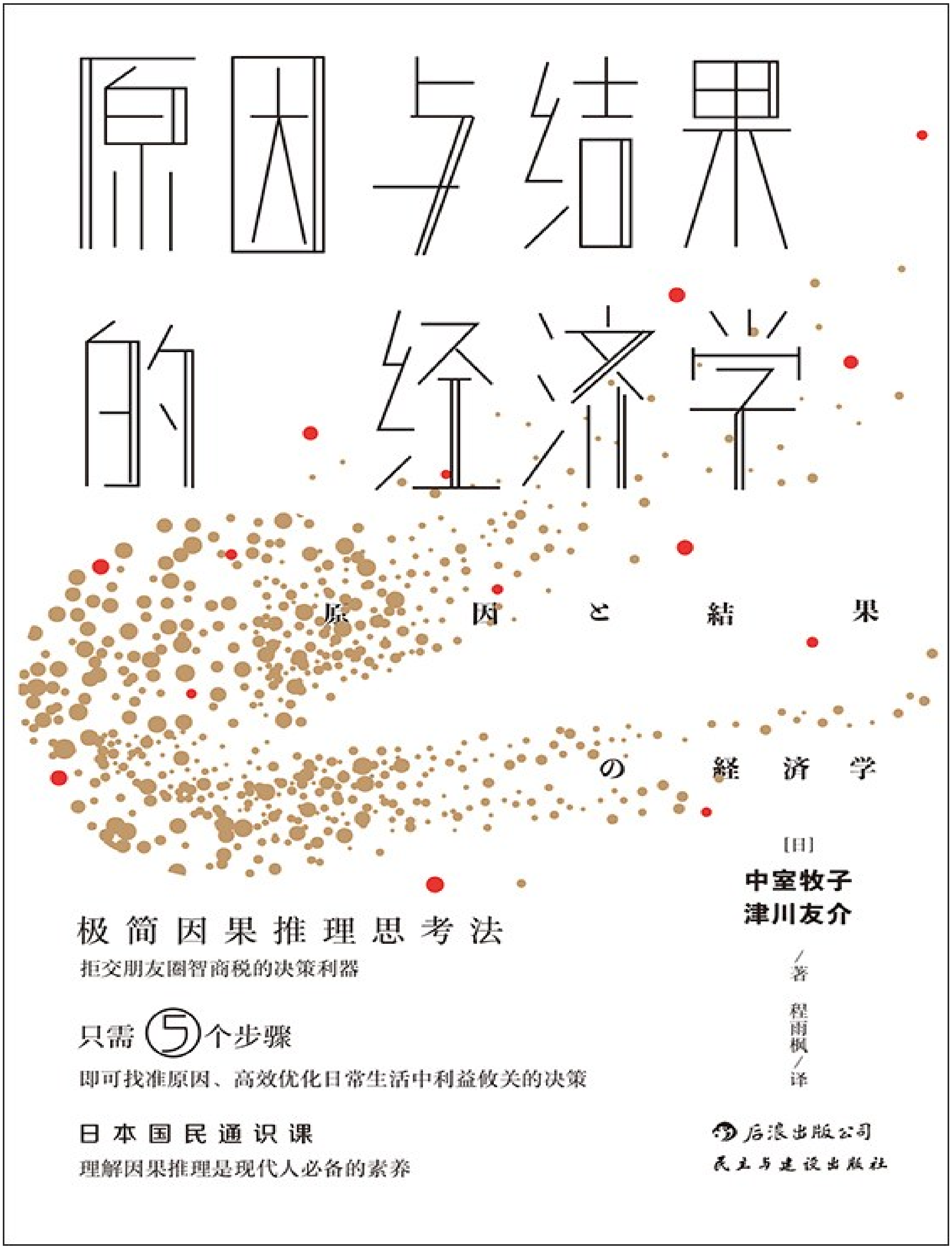
后浪出版公司

VISIT…

LANZAROTE Caliente.com

总目录

原因与结果的经济学
简单统计学
如何用数据解决实际问题
人生新算法：用人工智能解读时间、幸运与财富


原因与结果的经济学
原因と結果の経済学
[日]
中室牧子
津川友介
/著
程雨枫/译
极简因果推理思考法
拒交朋友圈智商税的决策利器
只需5个步骤
即可找准原因、高效优化日常生活中利益攸关的决策
日本国民通识课
理解因果推理是现代人必备的素养
后浪出版公司
民主与建设出版社

原因与结果的经济学

[日]中室牧子，[日]津川友介　著
程雨枫　译

版权信息

书名：原因与结果的经济学
作者：[日]中室牧子，[日]津川友介
译者：程雨枫
书号：978-7-5139-2295-1
版权：后浪出版咨询（北京）有限责任公司

前 言

在本书的开篇，我先请各位读者回答以下几个问题。

- “定期接受代谢综合征体检就能长寿吗？”
- “看电视会导致孩子学习能力下降吗？”
- “上偏差值①高的大学收入就会更高吗？”

想必很多人的回答都是肯定的。

不过，经济学的权威研究已经推翻了上述全部说法。

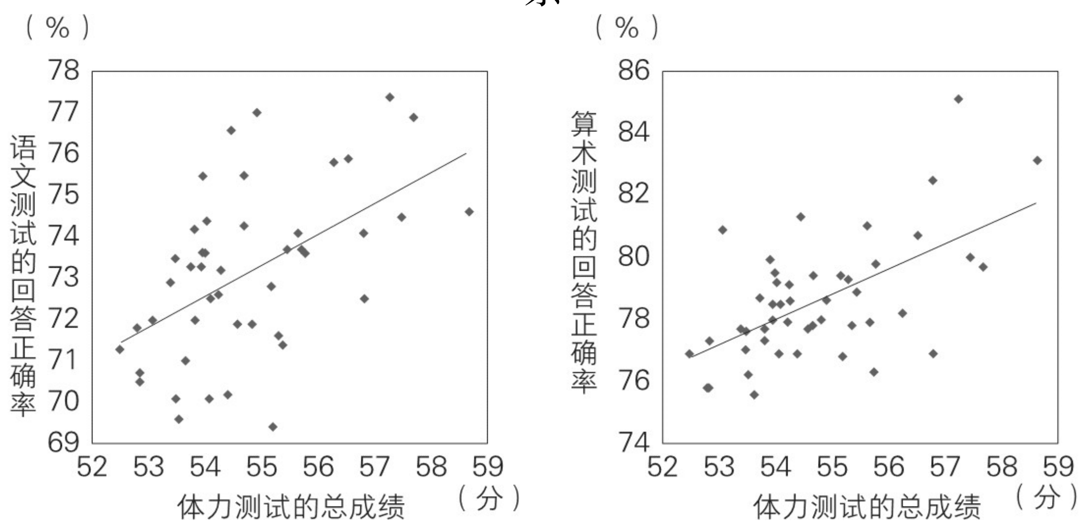
很多人之所以会做出肯定回答，是因为他们混淆了“因果关系”和“相关关系”。建议那些不假思索做出肯定回答的人务必读一读本书。他们读后定会收获全新的重要发现。

我们来看下面这个例子。

人们常说体力好的孩子往往学习能力也更强。图表1体现了日本各行政区孩子的体力测试与学力测试平均值的关系。从图表中可见，学生体力测试分数高的行政区，学力测试的正确率也更高。

那么，我们可否得出结论，认为他们“因为体力好，所以学习能力强呢”？换句话说，要想提高孩子的学习能力，是否应该先增强孩子的体力呢？

图表1　体力好的孩子学习能力强？——小学生的体力和学习能力的关系



（注）该图表体现了日本各行政区体力测试和学力测试的平均值的关系。由笔者根据《2014年度全国体力、运动能力及运动习惯等调查》（文部科学省）和《2014年度全国学力及学习情况调查》（国立教育政策研究所）的数据制作而成。

当然没有这回事儿。在经济学领域，“两个事件中，一方为原因，另一方为结果”的状态被称为存在因果关系。如果体力好这项“原因”会产生学习能力强的“结果”，便可以说二者之间存在因果关系。另一方面，“两个事件相互关联，但不存在因果关系”的状态被称为相关关系。在相关关系中，即使“看似是原因的事件”再次出现，也得不到期望中的“结果”。可见分清因果关系和相关关系是十分重要的。

“因为体力好，所以学习能力强”，其言下之意是“只要增强体力，（即便整天不学习）学习能力就会有所提高”。这种说法显然有问题，因此体力和学习能力的关系不是因果关系，而是相关关系。当然，即使增强孩子的体力，孩子的学习能力也不一定会提升。

这个事例教给我们一个非常重要的教训：混淆因果关系和相关关系，会导致错误判断。

在关于儿童的体力和学习能力的这个事例中，大多数人应该都能分清因果关系和相关关系。不过我们的生活中充斥着五花八门的信息，其中不乏听起来很有道理的说法。遗憾的是，很多人面对类似信息，都会不假思索地把相关关系视为因果关系。现在，让我们再回过头来思考本书开篇提出的三个问题吧。

- “定期接受代谢综合征体检就能长寿吗？”
- “看电视会导致孩子学习能力下降吗？”
- “上偏差值高的大学收入就会更高吗？”

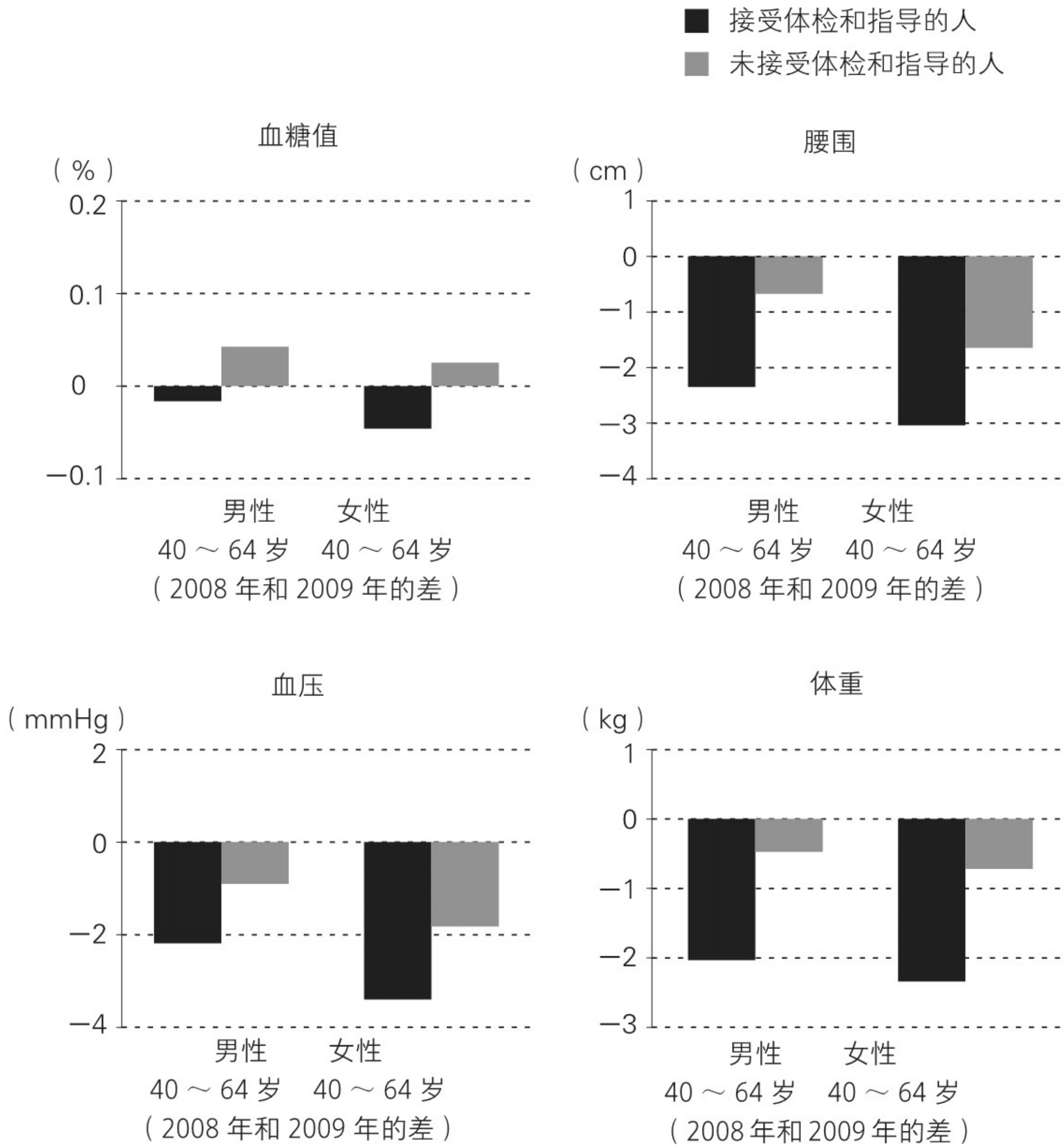
要让这三个说法成立，“代谢综合征体检”和“长寿”、“看电视”和“孩子的学习能力”、“大学的偏差值”和“收入”之间必须存在因果关系，而非相关关系。那么，每组事件之间的关系究竟是因果关系还是单纯的相关关系呢？下面让我们来依次探讨。

定期接受代谢综合征体检就能长寿吗?

学校或公司经常为我们安排体检[2]，其中最常见的是“代谢综合征体检”。想必很多人相信，通过代谢综合征体检掌握自己的健康状况，预防生活习惯病，或者发现潜在疾病，我们就会更长寿。[3]

图表2的柱状图显示，在代谢综合征体检中接受了生活习惯指导的人在第二年出现了腰围变小、体重下降、血糖或血压降低等变化。

图表2　定期接受代谢综合征体检的人身体健康吗？——代谢综合征体检与健康的关系



（注）代谢综合征体检结果为生活习惯病患病风险高的人提供保健指导。指导包括动机支援（原则上只进行1次）和积极支援（持续定期指导）。本图表中将参加积极支援的人称为"接受体检和指导的人"，未参加积极支援的人称为"未接受体检和指导的人"。血糖值为糖化血红蛋白（HbA1c），血压为收缩压数值。

（数据来源）摘自第14次保险人体检及保健指导研讨会的资料（厚生劳动省2015年6月26日），有部分变更。

乍一看，似乎确实可以说接受代谢综合征体检就会更健康、更长寿。

不过，请不要急于得出结论。这份数据真的可以说明“因为定期接受了代谢综合征体检，所以长寿”吗？回答这一问题的关键在于弄清代谢综合征体检和健康是因果关系还是仅为相关关系。这份数据还可以如此解读：并非“因为定期接受了代谢综合征体检，所以长寿（因果关系）”，而只不过是“关注健康、定期接受代谢综合征体检的人更长寿”（相关关系）而已。

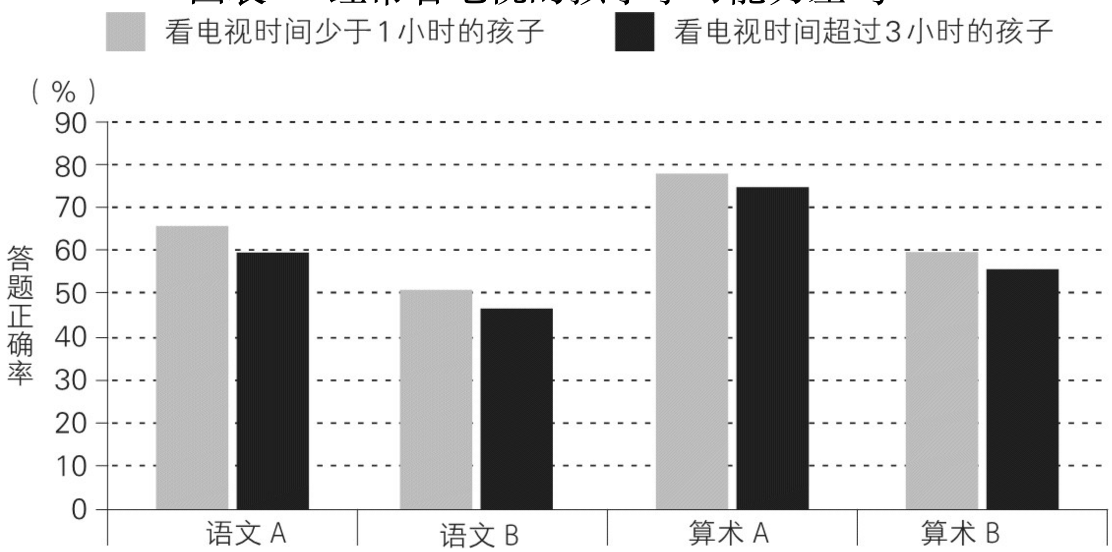
先来公布答案。经济学的权威研究证实，代谢综合征体检和长寿之间不存在因果关系。因此，“我每年都接受体检，所以身体很健康”是一种很危险的认知。详细讲解请见第2章。

看电视会导致孩子学习能力下降吗？

“我家孩子总是看电视”，想必不少家长都有这个烦恼。据日本厚生劳动省统计，小学六年级学生平日平均每天会在电视机前度过约2.2小时，周末平均每天在电视机前度过约2.4小时。

图表3也显示，每天看电视时间超过3小时的孩子比看电视时间少于1小时的孩子的学力测试成绩低。光从这个数据来看，似乎看电视确实会对孩子的学习能力造成不良影响。

图表3　经常看电视的孩子学习能力差吗?



（注）学力测试结果采用的是小学六年级学生的“全国学力学习情况调查”中语文和算术的答题正确率。A 为基础题，B 为应用题。

（数据来源）日本文部科学省《2015 年度全国学力调查》。

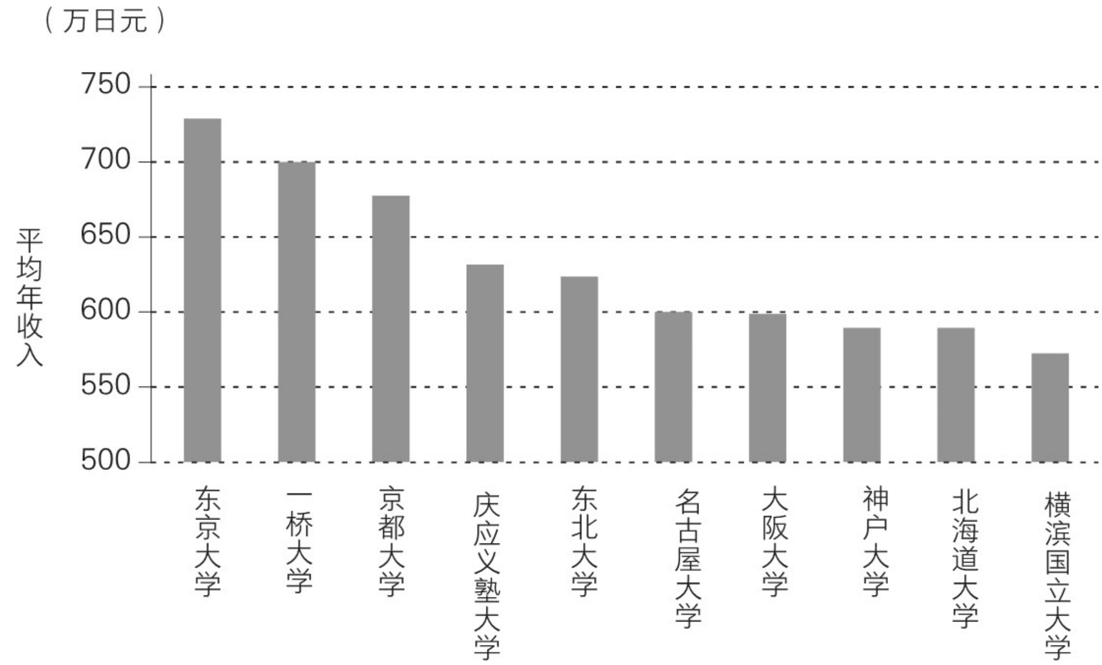
但是，我们必须认真思考，看电视和孩子学习能力之间的关系属于因果关系还是相关关系。也就是说，要分清是“因为看电视，所以学习能力下降”（因果关系），还是只是“学习能力差的孩子看电视时间更长”（相关关系）而已。

关于这个问题，我们也先公布答案。经济学的权威研究证实，看电视的时间和学习能力之间确实存在因果关系，但研究发现，看电视时间越长，学习能力不但不下降，反而提升。详细讲解请见第5章。

上偏差值高的大学收入就会更高吗?

很多人坚信“考上偏差值高的大学，未来就能拿高薪”。从实际数据中也能看出，高偏差值大学的毕业生具有年收入更高的倾向。（见图表4）

图表4　高偏差值大学的毕业生收入更高吗？——各大学毕业生的平均年收入（前10 位大学）



（数据来源）DODA 网：https://doda.jp/careercompass/ranking/daigaku_nenshu.html.

思考这个问题的关键依然是，大学偏差值和年收入之间是因果关系还是相关关系。或许不是“因为上了偏差值高的大学，所以收入高”（因果关系），而只是“有潜力将来获得高薪的人才多集中在偏差值高的大学”（相关关系）。

关于这个问题，我们也先公布答案。经济学的权威研究证实，大学偏差值和未来的年收入之间不存在因果关系。上偏差值高的大学并不代表步入了人生的康庄大道。详细讲解请见第7章。

或许有人不明白，不存在因果关系又有什么影响。或许有人会认为：接受代谢综合征体检总好于不体检，控制看电视的时间总比抱着电视不放要强，而上偏差值高的大学总比上偏差值低的大学要好吧。

不过不要忘了，我们采取任何行动都需要花费可观的金钱和时间。偏信那些看似存在因果关系的“无稽之谈”，把它们作为行动依

据，不仅得不到预期的成效，还白白浪费了金钱和时间。如果依据因果关系有效地把这些金钱和时间利用起来，我们获得满意的成果的概率也会更高。

理解“因果推理”，便可摆脱传统观念的束缚

两个事件之间是否真的存在因果关系？为解答这一问题，近年来经济学研究倾注了巨大的心血。我们把正确区分因果关系和相关关系的方法论称为“因果推理”。

“因果”，顾名思义，即“原因与结果”。“推理”则指“根据某个事件推导其他事件，即经过推测和推断得出结论的过程”。换句话说，就是分析并判断两个事件是否分别为原因和结果的过程。

在日常生活中，只要理解了因果关系和相关关系的差异，训练自己思考是否存在因果关系，便可以摆脱偏见或无稽之谈的束缚，做出正确的判断。本书的目的便在于用浅显易懂的方式详尽讲解因果推理最根本的思维方法。读者可以把本书当作因果推理的“入门的入门”。既然是入门的入门，自然不涉及经济学的基础知识，更不会用到任何公式。

此外，本书还将用大量篇幅介绍经济学中运用因果推理与数据得出的研究成果，说明如何解读这些研究。当今社会，“大数据”已成为脍炙人口的词汇，任何人都能轻而易举地分析数据，但是这并不代表所有人都能正确解读数据分析的结果。若想在大数据时代获得一席之地，不仅要学会分析数据，更要掌握解读数据分析结果的能力。

在此请允许我们做个自我介绍。本书作者之一的中室牧子是教育经济学家，多年来一直致力于运用数据和经济学方法研究何种教育方法能提高儿童的学力及能力。她所提倡的不是基于个人经验的教育论，而是呼吁根据具有因果关系的科学依据来制定育儿及教育政策。

另一名作者津川友介是医生兼医疗政策专家，致力于运用大数据研究如何在改善医疗质量的同时抑制医疗费用的增加。津川师从美国著名医疗经济学家、任教于哈佛大学的约瑟夫·纽豪斯（Joseph Newhouse）以及最早创建因果推理体系的唐纳德·鲁宾（Donald B. Rubin），从他们那里学会了因果推理的思维方法。

在美国，“因果推理”是大学的课程内容之一。商务和政治场合自不必说，日常会话中很多人也会有意识地在言谈中注意因果推理的应用。而在日本却几乎没有系统学习因果推理的机会。或许因为这个原因，电视和报纸上常有报道将一些相关关系描述成因果关系，甚至企业管理者和政策制定负责人也常把因果关系和相关关系混为一谈。

教育领域和医疗领域里充斥着无数因果关系混淆的无稽之谈。本书以生活中必不可少的教育和医疗为例，尽可能让各位读者在阅读本书后掌握因果推理的基本思维方法。

19世纪杰出的美国思想家、作家拉尔夫·爱默生说过这样一句话：“浅薄的人相信运气，强大的人相信因果。”在这个数据泛滥的时代，“因果推理”是我们每个人必备的素养。

① 日本普遍使用的用来评定学生考试成绩的方法。偏差值=（得分-平均分）÷标准差×10＋50。

② “体检”可分为“健康检查”和“疾病筛查”两种。前者旨在预防代谢综合征等生活习惯病，后者旨在及早发现特定疾病，如癌症筛查等。此处的“体检”指的是“健康检查”。

③ 代谢综合征体检的指标包括腰围、中性脂肪、高密度脂蛋白胆固醇、血压和空腹血糖。也有观点质疑这些诊断指标缺乏科学依据。

目　录

1. 前　言
2. 第1章　如何不受无稽之谈的蒙蔽 因果推理最根本的思考法
    1. 何谓“因果关系”和“相关关系”？
    2. 判断因果关系的三个要点
    3. 证明因果关系需要“反事实”
    4. 没有时光机就制造不出反事实吗？
    5. 用“最贴切的值”替换反事实
    6. 只有“可比较”的组才能替换
    7. 想象不出正确的反事实就会被无稽之谈蒙蔽吗？
    8. COLUMN 1 巧克力消耗量越大，诺贝尔奖获奖人数越多？
3. 第2章　定期接受代谢综合征体检就能长寿吗？ 因果推理的理想形态——随机对照试验
    1. “实验”能证明因果关系

2. 随机分组的必要性

3. “代谢综合征体检”与“长寿”是因果关系吗？

4. 何谓“在统计学上具有显著性”？

5. 定期接受体检并不能带来长寿

6. 投入1 200亿日元税金的代谢综合征体检

7. “医疗费用自付比例”和“健康”是因果关系吗？

8. 兰德健康保险实验的结果

9. 除贫困阶层以外，提高自付比例对健康状况没有影响

10. COLUMN 2 整合多项研究的“元分析”

4. 第3章　男医生比女医生更优秀吗？　利用与实验类似的偶发现象进行自然实验

1. 用现有数据重现与实验类似的环境

2. “医生性别”与“患者死亡率”是因果关系吗？

3. 女医生负责的患者死亡率更低

4. “出生体重”与“健康”是因果关系吗？

5. 出生体重较重的婴儿更健康

6. COLUMN 3 被动吸烟会增加心脏病的患病风险吗？

5. 第4章　最低工资与就业之间存在因果关系吗？　排除趋势影响的双重差分法

1. 模仿实验的“准实验”
2. 前后比较毫无意义
3. 不能使用实验前后测分析的两个原因
4. 去年的销售额为“反事实”时，前后比较才有效
5. 实验前后测设计的改良版——双重差分法
6. 双重差分法成立的两个前提条件
7. “保育所数量”与“母亲就业”是因果关系吗？
8. 增设认可保育所不会提升母亲的就业率
9. “最低工资”与“就业”是因果关系吗？
10. 提高最低工资不会减少就业
11. COLUMN 4 “不快点睡觉，妖怪就要来了”是正确的教育方法吗？

6. 第5章　看电视会导致孩子学习能力下降吗？　利用第三变量的工具变量法
    1. 广告费打折，该怎样利用
    2. 工具变量法成立的两个前提条件
    3. “看电视”与“学习能力”是因果关系吗？
    4. 看电视可以提高偏差值
    5. “母亲的学历”与“孩子的健康”是因果关系吗？

6. 母亲上过大学，孩子更健康
7. COLUMN 5 增加女性管理层成员能促进企业成长吗？

7. 第6章　和学霸做朋友，学习能力会提高吗？ 关注跳跃的断点回归设计
    1. 利用“49人店铺”与“50人店铺”的差异
    2. 断点回归设计成立的前提条件
    3. “朋友的学力”与“自己的学力”是因果关系吗？
    4. 和学霸在一起，也无法提高自己的学习能力
    5. “医疗费用自付比例”与“死亡率”是因果关系吗？
    6. 提高老年人医疗费用的自付比例对死亡率没有影响
    7. COLUMN 6 “激素替代疗法”的陷阱
8. 第7章　上好大学收入就会更高吗？ 组合相似个体的匹配法
    1. 找出相似的店铺
    2. 整合多个协变量的“倾向得分匹配法”
    3. 倾向得分匹配法成立的前提条件
    4. “大学偏差值”与“收入”是因果关系吗？
    5. 上偏差值高的大学，收入也不会增长
    6. COLUMN 7 商务版随机对照试验“A/B测试”
9. 第8章　便于分析现有数据的回归分析

1. 假如现有数据都不适合用来评估因果关系……
2. 画一条表示数据的"回归线"
3. 用"多元回归分析"排除混杂因素的影响
4. COLUMN 8 因果推理的发展史

10. 补论1　了解分析的有效性和局限性
 1. 内部效度与外部效度
 2. 随机对照试验也有局限性
11. 补论2　因果推理的五个步骤
12. 后　记
13. 参考文献

返回总目录

第1章　如何不受无稽之谈的蒙蔽
因果推理最根本的思考法

何谓“因果关系”和“相关关系”?

让我们再来看看因果关系和相关关系的概念。

如果两个事件中，前一个事件是后一个事件的原因，后一个事件是前一个事件的结果，则两个事件之间存在“因果关系”；如果一个事件变化后，另一个事件也随之发生变化，但二者不属于原因和结果的关系，则称它们之间存在“相关关系”①。（见图表1-1）存在相关关系的两个事件之间虽然有关联，但不属于因果关系。

图表1-1　因果关系与相关关系



这里所说的“事件”等可以取各种值的数据称为“变量”。变量可以是年龄、身高等数值，也可以如性别等取男性或者女性等文字作为值。本书在论述时将变量分为“原因”和“结果”②两大类。

两个变量的关系是否真的是因果关系？解答这个问题所需的思维方法便是“因果推理”。

判断因果关系的三个要点

判断两个变量属于因果关系还是相关关系时，可以通过以下三个问题进行质疑。

①是否“纯属巧合”？

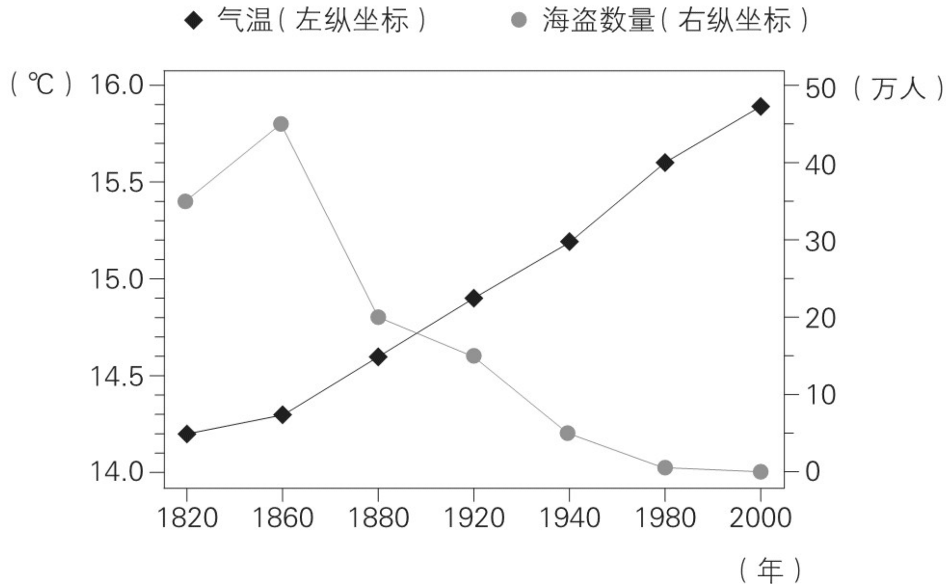
②是否存在“第三变量”？

③是否存在“逆向因果关系”？

①是否“纯属巧合”？

如果有人提出“温室效应的加剧导致海盗数量下降”，大家肯定觉得荒唐，但如图表1-2所示，温室效应逐渐加剧的同时，海盗数量确实在不断减少。

图表1-2　温室效应加剧会导致海盗数量下降吗？——纯属巧合



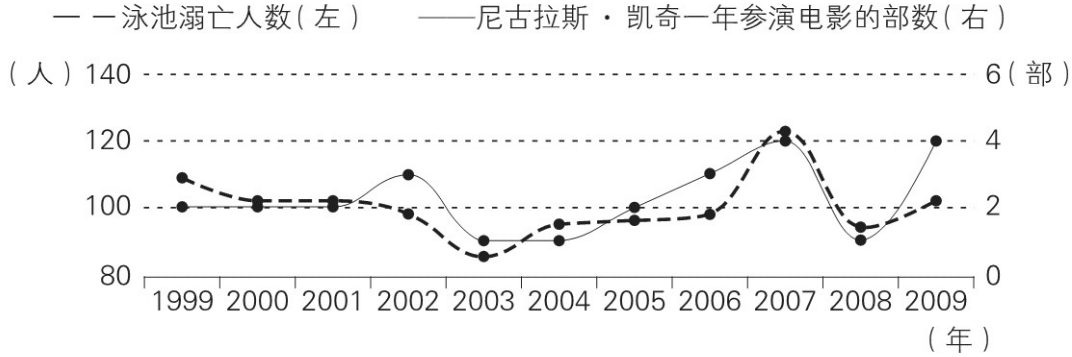
（数据来源）本图表由笔者根据福布斯网站（http://www.forbes.com/sites/rikaandersen/2012/03/23/true-fact-the-lack-of-pirates-is-causing-global-warming/#1606f14ca231）的数据制作而成。

可是，按常理来说，很难想象温室效应的加剧会导致海盗数量下降。两个事件看似有关，其实只是“纯属巧合”。像这样，两个变量的变化趋势只是碰巧相似的现象被称为“伪相关”。

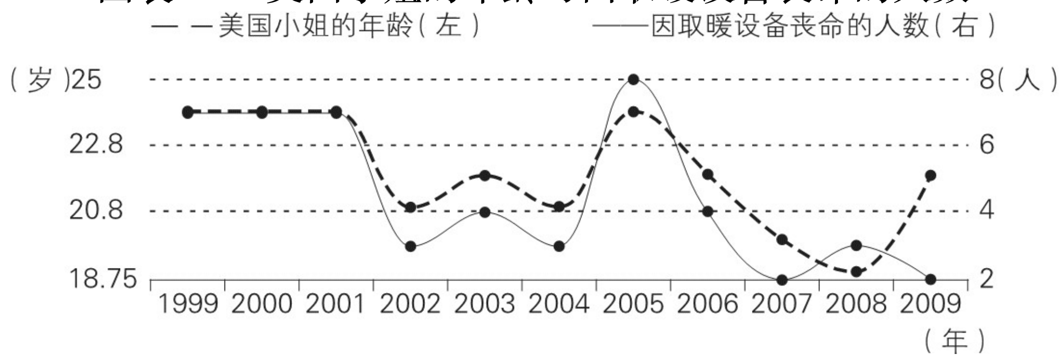
美国信息分析员泰勒·维根（Tyler Vigen）在其著作《伪相关》（*Spurious Correlations*）中介绍了很多“纯属巧合”的事例。例如，“尼古拉斯·凯奇一年参演电影的部数”与“泳池溺亡人数”（见图表1-3）、“美国小姐的年龄”与“因取暖设备丧命的人数”（见图表1-4）、“商

业街的总收入”与“在美国获得计算机科学博士学位的人数”（见图表1-5）等之间都存在高度的相关关系。

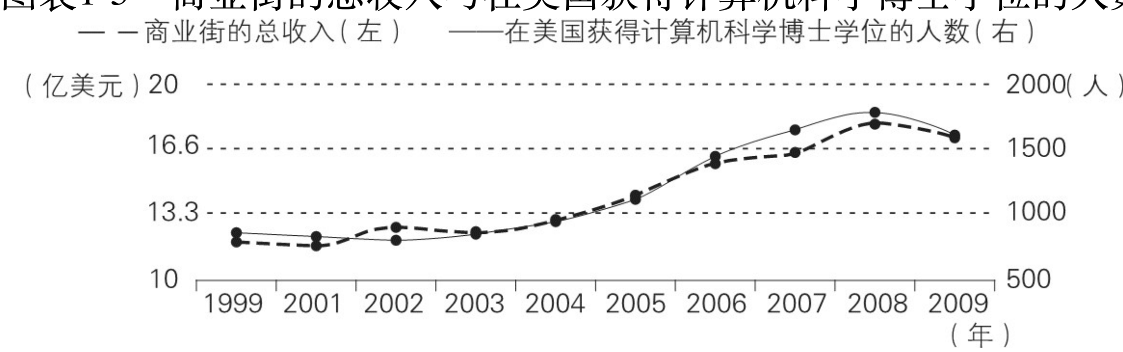
图表1-3　尼古拉斯·凯奇一年参演电影的部数与泳池溺亡人数



图表1-4　美国小姐的年龄与因取暖设备丧命的人数



图表1-5　商业街的总收入与在美国获得计算机科学博士学位的人数



（数据来源）泰勒·维根网站：http://tylervigen.com/spurious-correlations.

这些例子看似非常荒唐，但不可思议的是，把两个变量做成图表比对一下就会显示出明显的相关关系。正好比“蝴蝶在亚洲扇动翅膀，南美就会刮起飓风”，不过需要留意的是，我们身边其实存在很多这种“纯属巧合”的相关关系 。

或许你会怀疑，真的会有人把伪相关当作因果关系吗？实际上，不少预测大盘的人都会相信纯属巧合的伪相关，把它们当作没有依据却非常灵验的经验之谈。

例如可能有人听说过“吉卜力的诅咒”。即只要日本的电视台播出宫崎骏导演领导的吉卜力工作室的电影，美国股市就会下跌。就连美国的《华尔街时报》也曾报道过这条法则并一度引起热议。这条法则正是“纯属巧合”造就的典型的伪相关③。

探讨“是否存在因果关系”时，一定要先质疑两个变量之间的关系是否纯属巧合。

②是否存在“第三变量”?

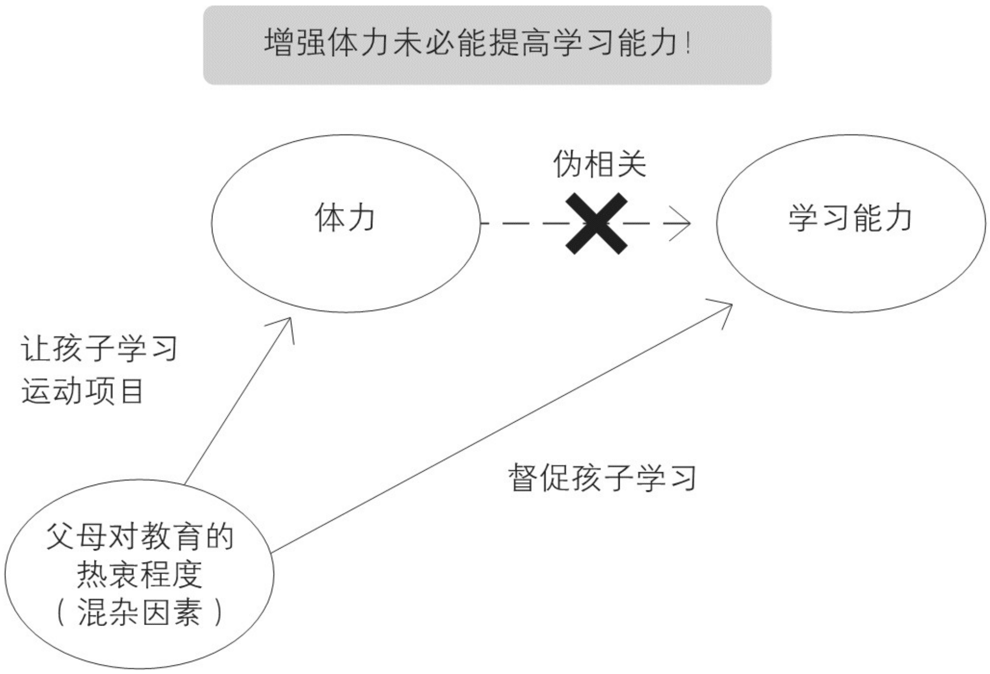
其次，我们必须要质疑是否存在同时影响原因和结果的“第三变量”。用术语说就是“混杂因素”④，它可以把纯粹的相关关系包装成因果关系，干扰人们判断。

我们来看一则混杂因素的具体事例。前言中曾经提到这个说法：体力好的孩子学习能力强。可能一些家长看到这种说法，就打算让孩子去锻炼身体。

然而，断定体力与学习能力之间存在因果关系还为时过早。说不定另有一个变量同时影响着孩子的体力和学习能力。（见图表1-6）比如“父母对教育的热衷程度”等。热衷教育的父母可能会让孩子学习运动项目或者注重饮食健康（对体力产生影响），同时还会督促孩子学习，因此孩子的学习能力也比较强（对学力产生影响）。这则事例中，提高

孩子学习能力的不是体力，而是“父母对教育的热衷程度”。如果事实果真如此，那么一味增强孩子的体力，恐怕不会提高他们的学习能力。

图表1-6　“混杂因素”把相关关系包装成因果关系



探讨“是否存在因果关系”时，切记还要质疑是否存在同时影响原因和结果的“混杂因素”。

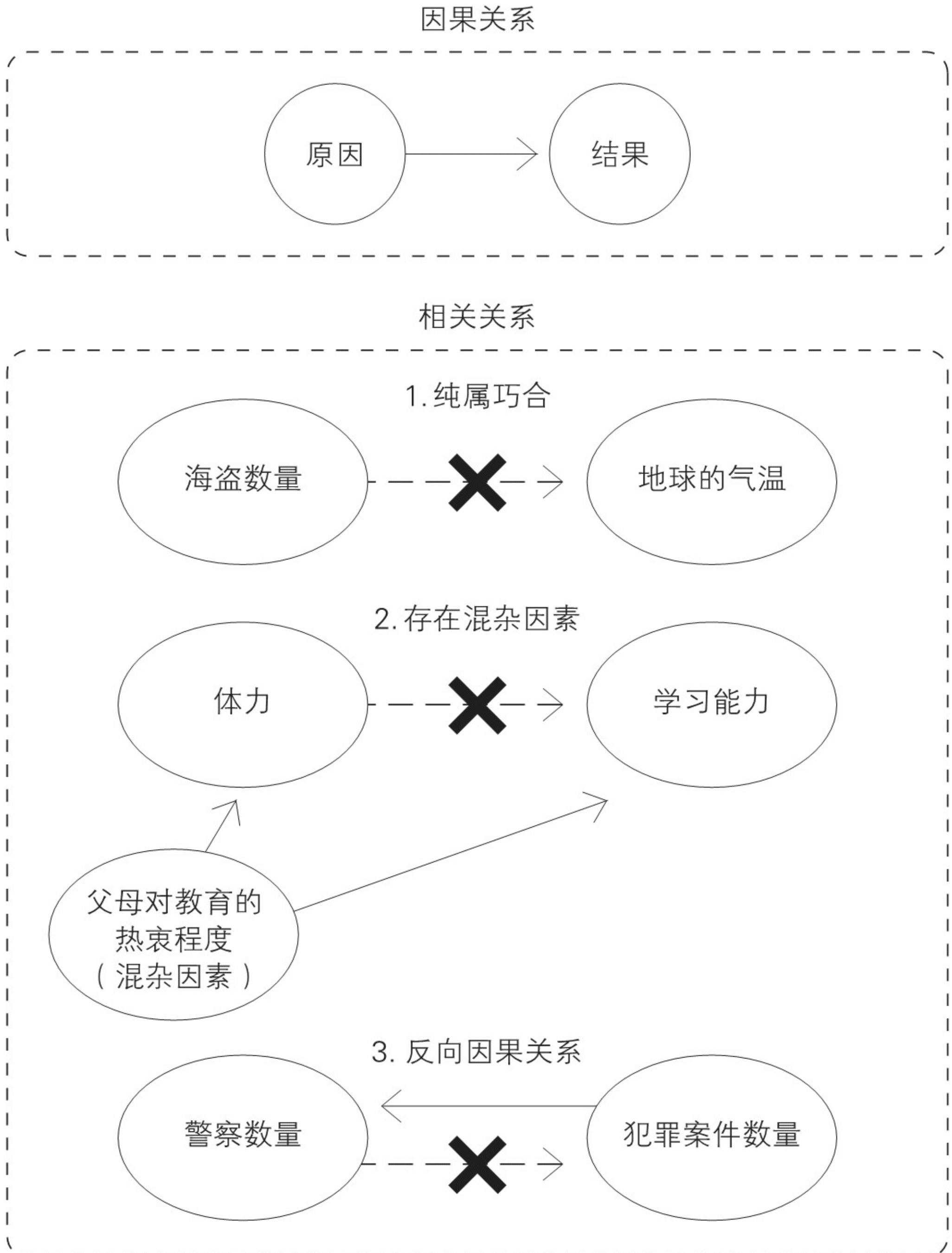
③是否存在“逆向因果关系”？

接下来需要质疑是否存在“逆向因果关系”。例如，我们来思考警察与犯罪的关系。警察多的地区，犯罪案件数量也多。但是，将警察多视为犯罪案件数量多的原因难免有些牵强（警察→犯罪）。

不如说因为某处是犯罪多发的危险地区，所以部署了大量警力，这么想可能才更合理（犯罪→警察）。本以为是原因的事件其实是结果，本以为是结果的事件其实却是原因，这种状态被称为“逆向因果关系”。探讨“是否存在因果关系”时，还要质疑原因与结果的方向是否相反。

让我们对照图表1-7对上述说明加以总结。

图表1-7　因果关系与相关关系的总结



如果两个变量之间存在因果关系，当原因再次出现时，相同的结果也会出现，而不存在“纯属巧合”“混杂因素”或“逆向因果关系”。另一方面，如果两个变量的关系只是相关关系，那么就会存在“纯属巧合”“混杂因素”或“逆向因果关系”中的某一种情况。在相关关系的情况下，即使原因再次发生，也几乎不会得到相同的结果。

证明因果关系需要“反事实”

那么，应该如何证明不存在上述三种情况呢？方法就是对现实和“反事实”⑤进行对比。反事实是指对过去未曾发生的事实所做的假设，例如“如果当时没有……，那么……”。我们将现实中实际发生的事称为“事实”，所以将设想的与现实完全相反的情况称为“反事实”。

在日常生活中，我们应该都有过这样的想法。

“如果当时没有跳槽到这家公司，我现在的收入会有多少呢？”

“如果当时下定决心和他结婚，现在我会过着什么样的生活呢？”

法国哲学家布莱士·帕斯卡曾经说过这样一句话：

“如果克莉奥帕特拉的鼻子再塌一点，世界史就会改写了。”

这正是典型的反事实思维。

要证明因果关系的存在，必须对“事实”中原因发生后的结果与“反事实”中原因未曾发生时的结果进行对比。

没有时光机就制造不出反事实吗?

“反事实”是因果推理中最为重要的概念，下面让我们举例详细介绍。

假设你在一家经营全国连锁珠宝店的企业担任公关部部长，正在考虑通过投放广告来提高销售额。所幸聘请当红小生制作的报纸广告十分具有吸引力，投放时期也定在12月上旬，刚好配合圣诞季。

结果如你所料，广告刊登后，客流量激增，销售额和去年同期相比增长了30%。店铺收到大量预订，全体员工都满负荷工作。看到这个场景，你肯定会感到十分欣喜，并会对社长这样说：

“今年的销售额比去年同期增长了30%。这都是（我策划的）报纸广告的成效！”

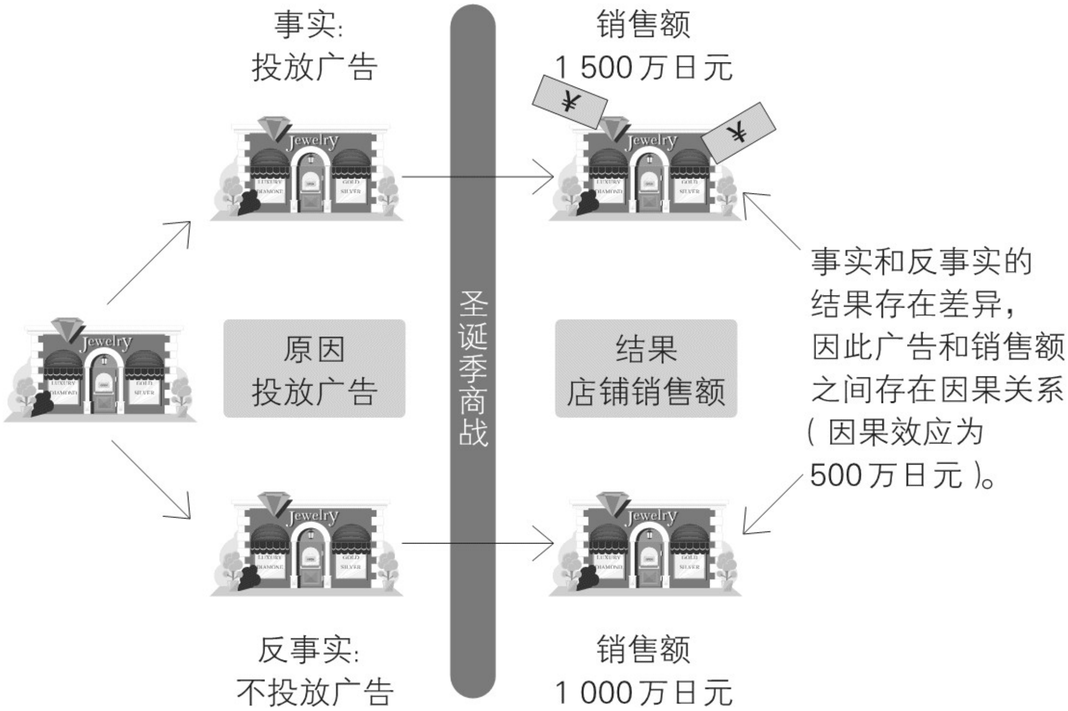
不过，不要着急，我们先来冷静地思考一下。广告和销售额之间真的存在因果关系吗？也许不是“因为投放了广告，所以引起销售额增长”（因果关系），说不定“即使不投放广告，销售额也会增长”（相关关系）呢。

那么，应该如何判断广告和销售额之间是否存在因果关系呢？

我们来做一个有些不切实际的想象。各位读者朋友，你们知道《回到未来》（*Back to the Future*）这部电影吗？在这部20世纪80年代大受好评的科幻电影中，科学家发明出了时光机。用这种时光机就能证明“因为投放了广告，所以销售额增长”的因果关系。

请看图表1-8。公司投放了广告。这种情况下，圣诞季商战的销售额为1 500万日元（事实中的销售额）。见证这一结果之后，你乘上时光机穿越到投放广告之前，让正打算投放广告的那个过去的自己打消了这个念头。这种情况下，圣诞季商战的销售额只有1 000万日元（反事实中的销售额）。

图表1-8　调查广告和销售额之间的因果关系



这样的话，就能肯定确实是因为投放了广告，所以销售额增长了，即广告与销售额之间存在因果关系。

也就是说，使用时光机，就能得知反事实“如果没有投放广告”状态下的销售额，明确广告与销售额之间是否存在因果关系。不仅如此，我们甚至还能掌握广告到底带来了多大的效果。用投放广告后的销售额1 500万日元减去未投放广告时的销售额1 000万日元，余下的500万日元便是广告带来的销售额。我们将其称为广告的“因果效应”⑥。

这里有一个问题：现实中我们可以观测事实，但无法观测到反事实。因为《回到未来》中的时光机如今依然是假想中的产物，我们不可能回到过去观测没有投放广告时的销售额是多少。

最早创建因果推理体系的哈佛大学统计学家唐纳德·鲁宾把这个问题称为“因果推理中的根本问题”。然而，要证明因果关系，反事实是必经之路。

用“最贴切的值”替换反事实

其实，克服“因果推理中的根本问题”，制造反事实，才正是以因果推理为基础的各种方法的根本所在。为此，经济学家会想方设法，用最贴切的值来替换无法观测到的“反事实的结果”。

让我们再回到珠宝店的事例。假设你所在的珠宝连锁店在全国四个地区设有店铺。为配合圣诞季商战，正打算投放广告。考虑到预算有限，最后决定在4个地区当中，为地区1和地区2的店铺投放广告。

假设你想了解广告能为销售额带来多大的因果效应。广告的因果效应等于投放广告后的事实销售额和相同情况下没有投放广告的反事实销售额的差。如果你能观测到反事实（虽然现实中不可能实现），因果效应将如图表1-9所示。

图表1-9　观测到反事实就能明确因果效应——虚构的珠宝店销售额

	地区	投放广告后的销售额（A）(万日元)	未投放广告时的销售额（B）(万日元)	因果效应（A-B）(万日元)
有广告	1	1 300	900	400
	2	1 700	1 100	600
无广告	3	1 600	1 200	400
	4	1 400	800	600
	平均值	1 500	1 000	**500**

图表1-9中的月销售额单位设为1万日元。图表中显示，地区1投放广告后的销售额为1 300万日元，而该地区未投放广告时的销售额为900万日元。

像这样，如果能掌握各地区投放广告和不投放广告时的销售额，就能计算出各地区广告的因果效应。对各地区的因果效应取平均值，就能得出广告为整个企业带来的因果效应。由图表1-9可见，广告对整个企业的因果效应为500万日元。

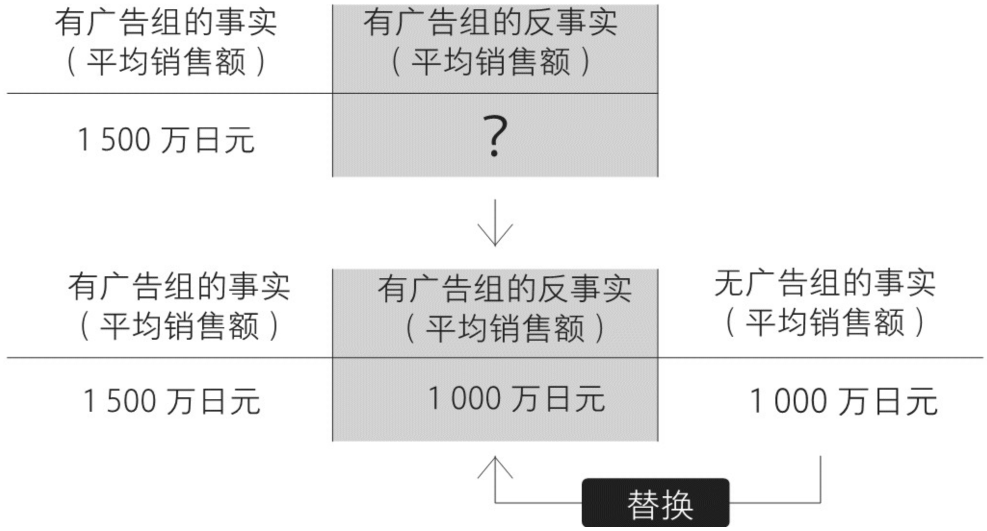
遗憾的是，现实中我们无法观测反事实中的销售额。实际能观测到的是图表1-10。这张表中的反事实销售额全部变成了“？”，除非拥有时光机，否则无法填写确切的数值。怎样把这个“？”替换为最贴切的数值呢？

图表1-10　现实中无法观测反事实——实际的珠宝店销售额

	地区	投放广告后的销售额（A）（万日元）	未投放广告时的销售额（B）（万日元）	因果效应（A–B）（万日元）
有广告	1	1 300	?	?
	2	1 700	?	?
无广告	3	?	1 200	?
	4	?	800	?
	平均值	1 500	1 000	?

我们可以换个角度思考，将投放广告的两个地区（地区1、地区2）和未投放广告的另外两个地区（地区3、地区4）分别归为一组。如此一来，有广告组未投放广告时的反事实销售额会不会和无广告组的销售额差不多？如果确实如此的话，只要用无广告组的销售额替换有广告组的反事实销售额就可以了。（见图表1-11）

图表1-11　用“最贴切的值”替换反事实



这样一来，广告对整个企业的因果效应就可以用有广告组的平均销售额（1 500万日元）减去无广告组的平均销售额（1 000万日元）求得，即500万日元（1 500万日元-1 000万日元）。

只有“可比较”的组才能替换

其实，只要满足一个条件，上述思路便可以成立。这个条件就是：有广告组和无广告组必须是“可比较”的。

假设有广告组位于大城市，当地居民收入很高。而无广告组位于地方城市，那里的居民一般不会购买宝石。这种情况下，就不能“用无广告组的销售额替换有广告组的反事实销售额”。

怎样才称得上“可比较”呢？如果两个组在人口、居民人均所得、流行敏感度等所有可能影响宝石饰品销售额的特征方面都非常接近，唯一的区别是“有无投放广告”，那么这两个组就是“可比较”的。

然而，现实中没有哪两个组在所有可能影响珠宝店销售额的特征上都非常接近。那么，也许有人会问：两个“大致相同”的组之间不能比较

吗?

很遗憾，“大致相同”不能和“可比较”画等号。让我们思考下面这则事例。两个组所在地区特征相似，除了有无投放广告之外，还有唯一一个小小的差异：投放广告组所在地区和未投放广告组所在地区播放的电视节目不同。当时投放广告的地区正在播放电视剧，参演的当红女演员佩戴了珠宝店的主打商品，而未投放广告的地区没有播放该电视剧。

即便只是很小的不同，我们也不应忽略由此带来的差异。这种情况下，即使有广告组的销售额高于无广告组，我们也无法判断带动销售额的是广告还是电视剧。假设电视剧的效应远远大于广告，会导致什么后果？第二年圣诞季即使再次投放广告，销售额恐怕也无法达到预期的水平。

因此，用于比较的两个组除了有无投放广告之外，在所有可能影响珠宝店销售额的特征上都必须极为相似。

然而，现实中我们很难找到这样的事例。正因为如此，经济学者们要运用各种方法，尝试将两个不相似的组转化为“可比较”的组。这些方法将在第2章之后详述。

想象不出正确的反事实就会被无稽之谈蒙蔽吗?

我想在这里反复强调的是，要明确是否存在因果关系，必须对比事实的结果和反事实的结果。不过遗憾的是，很多人尚未意识到这一点。一般来说，在未能得到正确的反事实时，人们常会把没有因果关系的事项错误地理解为存在因果关系。

例如，假设一位把孩子送入名牌大学的母亲在书中写道：“我从来不让孩子看电视。”那么读了这本书的很多人便会认为孩子学习能力高是因为家长没有让孩子看电视。同理，如果一位长寿的老人在百岁生日

当天接受电视台采访时说“我每年坚持体检”，观众便会以为“老人长寿是因为做了体检”。然而，要想证明看电视与学习能力、体检与长寿之间存在因果关系，必须用“同一个孩子在看电视的情况下的学习能力”“同一位老人在没有每年接受体检的情况下的寿命”等反事实的结果进行对比。

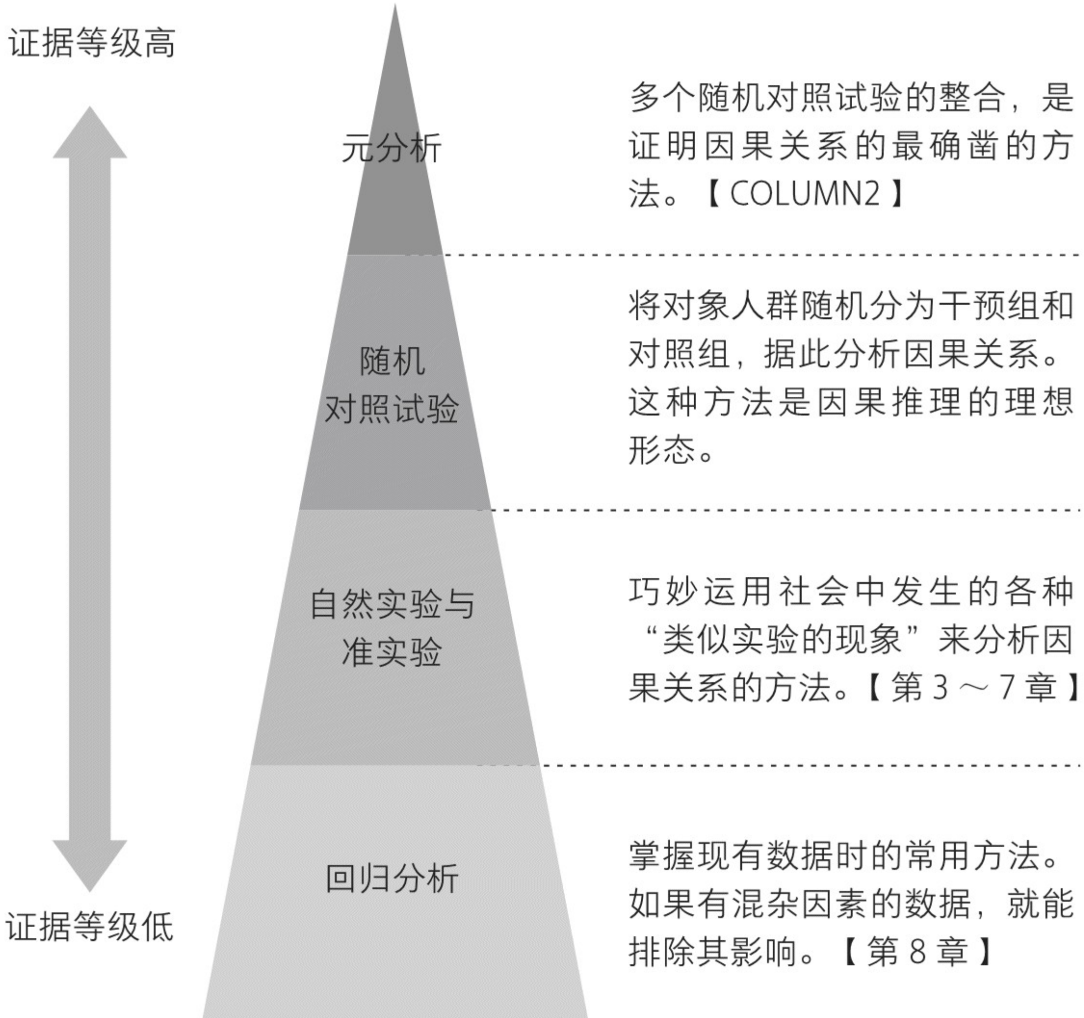
成功人士的故事里只包括事实，而不包括反事实。如果忽略了这一点，仅凭事实就片面地认为存在因果关系，一味禁止孩子看电视或者频繁接受体检，很可能不但无法成功，反而白白浪费了时间和金钱。

从第2章起，本书将具体介绍如何制造可比较的组，以及如何用最贴切的值替换反事实。在此我要重申的是，明确因果关系的方法有很多种，但是这些方法本质上的共同目标都是“制造可比较的组，用最贴切的值替换反事实”。在阅读本书时，请不要忘记下面介绍的所有方法都是为了实现这个目标。

接下来在介绍本书章节构成的同时，对“科学依据（证据，evidence）的等级”也稍做介绍。或许有人听说过“科学依据（证据）”一词。坊间有时将数据分析得出的根据称作科学依据，不过在经济学界使用这个词时要更加严谨：它指的是反映因果关系的证据。因此，经济学家不会把只证明了相关关系的分析称为证据，更不要说简单的图表或问卷调查结果了。

证据的概念在医学界应用得最为普及。医学领域一般认为，证据有“等级”之分。有些证据等级高，可信度高；有些证据等级低，可信度低。证据根据可信度分为多个层次，可信度高的证据是由能正确判断因果关系的方法推导出来的。与之相对，可信度低的证据是通过可能混淆因果关系和相关关系的方法推导出来的。图表1-12叫作“证据金字塔（evidence pyramid）”，越接近金字塔的顶点，证据的可信度越高。

图表1-12　证据的等级



* 也有一些元分析（meta-analysis）整合的是多个观测研究，其证据等级不如整合随机对照试验的元分析高。元分析的证据等级高低取决于其使用的研究的证据等级。

本图表由笔者根据萨基特（Sackett）等人的研究（2000）制作而成。

本书将按照证据金字塔的顺序逐一讲解。第2章介绍“随机对照试验”，第3章介绍“自然实验”，第4～7章介绍“准实验”的各种方法，第8章介绍“回归分析”。

下面就让我们运用这些方法，走近经济学发现的那些惊人的研究成果吧。

COLUMN 1
巧克力消耗量越大，诺贝尔奖获奖人数越多?

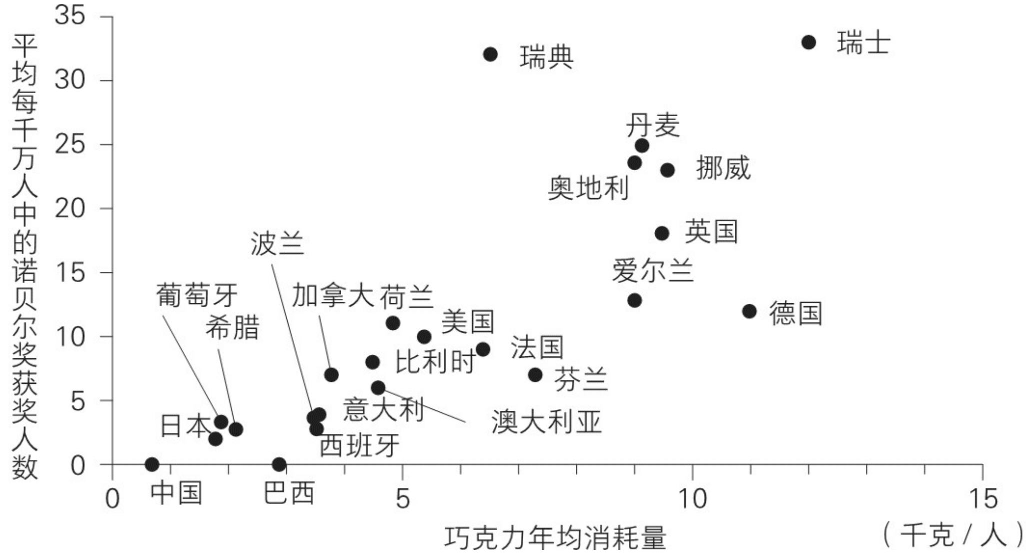
巧克力中的黄酮醇能够提高认知功能，这一结论早在动物实验等研究中得到证实。哥伦比亚大学的医师在2012年针对该结论进行数据分析，发现巧克力人均年消耗量越大的国家，诺贝尔奖获奖人数越多。该研究成果被刊登在临床医学界最具权威性的期刊之一《新英格兰医学杂志》上，引发了广泛议论。

论文作者得出的结论是：如果平均每人每年多摄入400克巧克力，该国的诺贝尔奖获得者就会增加一人。这种说法看似荒诞，但毕竟是刊登在权威期刊上的论文，其结论是否可信呢?

让我们回想起前言中的思路。“巧克力人均年消耗量”与“诺贝尔奖获奖人数”的关系是因果关系还是相关关系？或许不是“因为巧克力消耗量大，所以获得诺贝尔奖的人多”（因果关系），而只是“诺贝尔奖获奖人数多的国家巧克力消耗量更大”（相关关系）。

如论文中也有收录的图表1-13所示，欧洲人均GDP较高的国家都集中在右上部分。巧克力不是生存必需品，而是所谓的奢侈品，所以富裕国家的摄入量自然更大。而同时，富裕国家有能力在教育上投入更多资金，产生诺贝尔奖获奖者的可能性也会增加。也就是说，巧克力消耗量和诺贝尔奖获奖人数的关系很可能不是因果关系，而是相关关系。

图表1-13　巧克力消耗量与诺贝尔奖获奖人数的关系



（数据来源）梅瑟利（Messerli）（2012）

要维护《新英格兰医学杂志》的立场的话，我们可以注意到这篇文章不是研究论文，而是刊登在论述个人观点的“不定期笔记”专栏上。这个专栏旨在“介绍一些个人经验及医学研究领域通常不涉及的事项”。

其实，最近还有实验对摄入黄酮醇之后的大脑进行了MRI（核磁共振）影像诊断和记忆测试。其结果显示，与几乎从不摄入黄酮醇的老年人相比，摄入黄酮醇较多的老年人在大脑功能和记忆力方面更好。虽然多吃巧克力就更有可能获得诺贝尔奖的说法言过其实，但摄入巧克力至少具有增强记忆力的功效。

① 广义的相关关系既包括“非因果关系”，也包括“因果关系”。本书为避免混乱，采用更加严谨（狭义）的定义。

② 在其他学术领域，“原因”又被称为“自变量”“解释变量”“处理”或“暴露变量”；“结果”又被称为“因变量”“被解释变量”或“结果变量”等。

③ 一旦被人们广为接受，就会出现刻意利用这一法则的投资行为，使其不再是纯粹的巧合。经济学将这种情况称为存在“噪音交易者”。

④ 经济学中的术语“遗漏变量”和“混杂因素”概念相似。术语的详细释义请参考专业书籍。

⑤ 即“反事实推理”，又称“反事实思维”。

⑥ “因果效应”有时又被称为“处理效应”。

第2章　定期接受代谢综合征体检就能长寿吗?因果推理的理想形态——随机对照试验

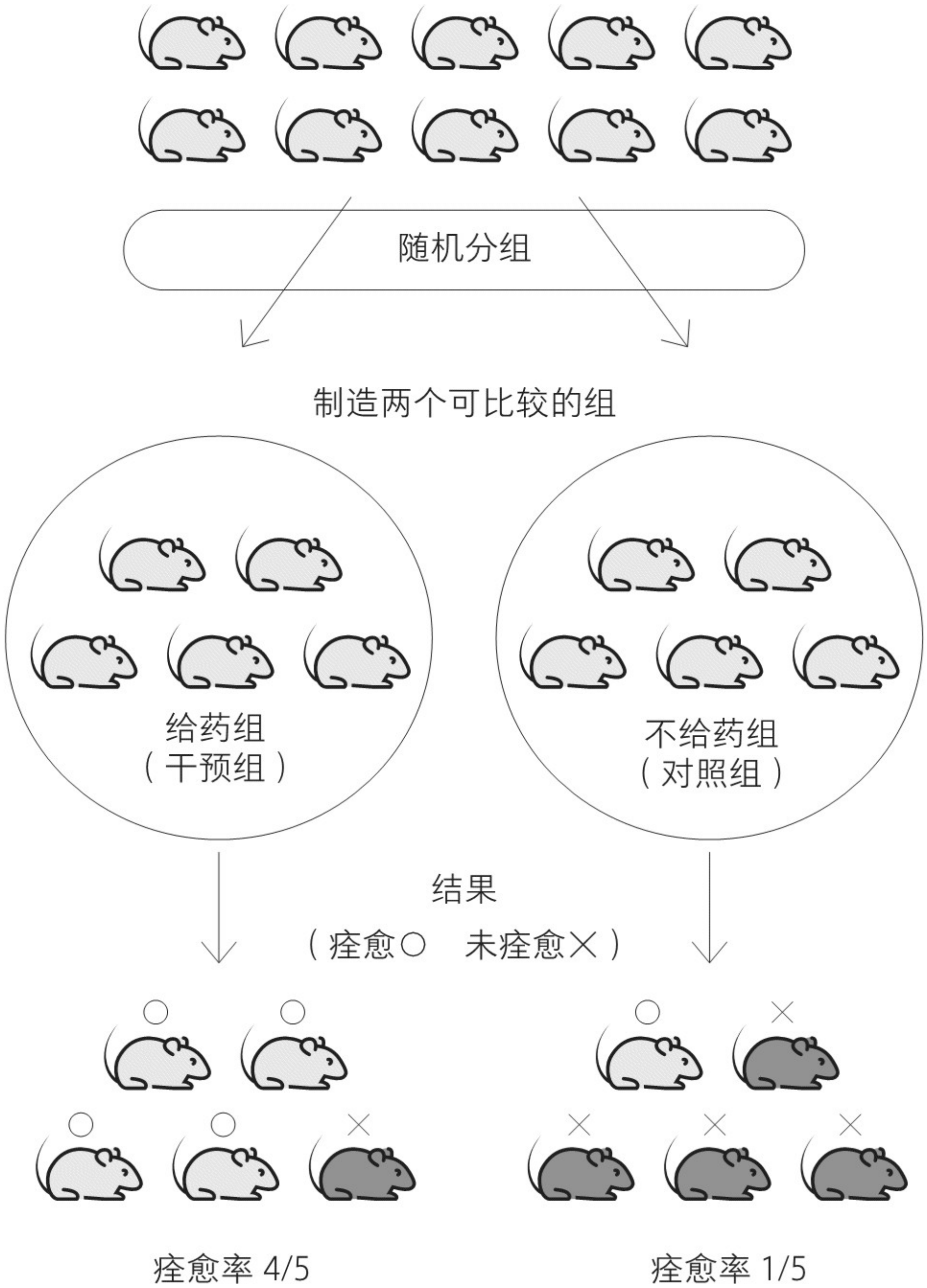
“实验”能证明因果关系

两个变量的关系属于因果关系还是相关关系？判断这个问题的最可靠方法是“实验”，专业术语称为“随机对照试验”。

各位读者朋友可能都听说过“临床试验”。这是为了明确药物疗效与安全性所进行的试验。

例如，在小鼠实验中，将生病的小鼠“随机”分成两组，对比给予药物的小鼠（称为干预组）和不给予药物的小鼠（称为对照组）。如果摄入药物的小鼠痊愈率更高，则可认为药物有效。（见图表2-1）

图表2-1　随机分组制造反事实



“随机”或许给人一种“随随便便”的感觉，其实不然。把实验小鼠划分为干预组或对照组时，所有小鼠都有相同概率被划入干预组的分组方法被称为“随机分组”。随机还带有“依赖偶然性”的含义。

实际操作中如何进行随机分组？例如可以投硬币，正面朝上的划入给药组，背面朝上的划入不给药组（反之亦可）；或者使用由随机生成的数字组成的随机数表，根据抽取的是奇数或偶数来决定是否给药；也可以用抽签的方式决定分组。

小鼠之间的个体差异不明显，只要准备足够多的个体进行随机划分，就不会受到小鼠个体差异所导致的偶然因素的影响。因此，给予药物的小鼠组和不给予药物的小鼠组可以视为可比较的两个组。

换句话说，随机对照试验的本质就是用对照组替换“如果干预组没有投放药物”的反事实。

随机分组的必要性

通过随机对照试验，便可以解答前言中提到的问题：定期接受代谢综合征体检就能长寿吗。

这种方法将研究对象人群随机分为两组：接受体检“干预”的组（干预组）和不接受体检的组（对照组）。这里的“干预”意为将原因与结果中相当于“原因”的事项（此处指体检）施加给研究对象人群。

为什么一定要进行随机分组？不能直接比较曾经接受过体检和未接受过体检的人吗?

其实曾经接受过体检和未接受过体检的人属于不可比较的群体。试想过去一直接受体检的人健康意识肯定很高，而未接受过体检的人恐怕都不太注重身体健康。所以不管怎样，这两类人群都不适合拿来做比较。

人与临床试验小鼠的不同之处在于，人会按照自身的意志选择自己的行为。个人“选择（selection）”会导致两个研究对象组不具备可比

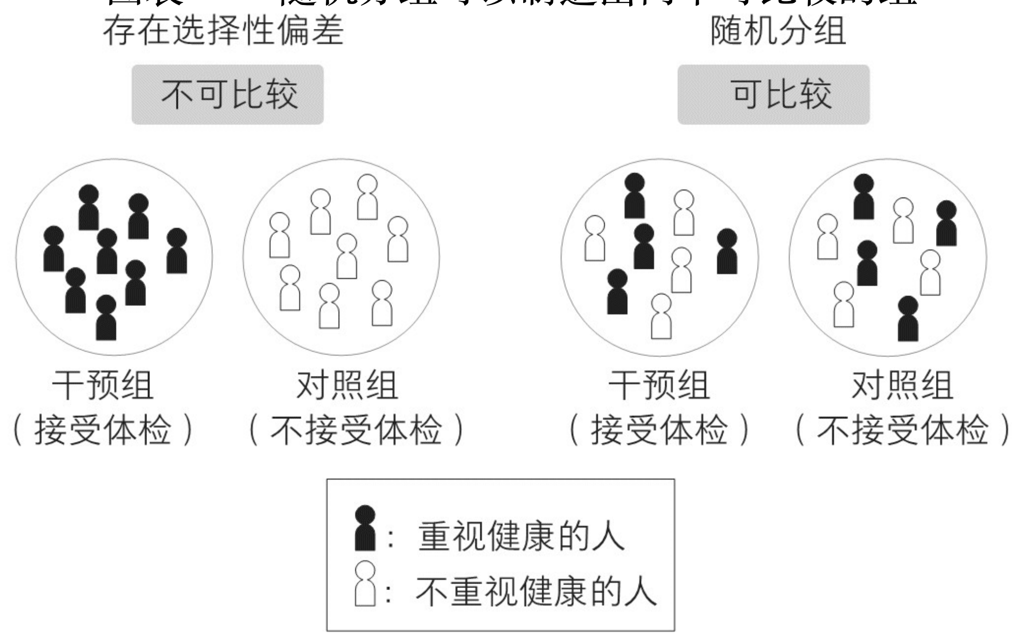
较的属性。这种现象在经济学中被称为“选择性偏差（selection bias）”。

英语中有这样一句俗语：简直是在拿苹果和橘子比。这句话讽刺了拿两种原本就截然不同、没有比较意义的事物做比较的行为。对曾经接受过体检的人和未接受过体检的人进行比较，就类似于“拿苹果和橘子比”。

那么，如何才能像拿苹果和苹果比（或者拿橘子和橘子比）一样，对曾经接受过体检的人和未接受过体检的人进行比较呢？

最可靠的方法就是用抽签等方式随机决定是否接受体检。（见图表2-2）通过随机分组，可以使个人无法按照自身的意志做出选择，因此也就不会出现选择性偏差。由此得出的接受体检组和不接受体检组便是可比较的。

图表2-2　随机分组可以制造出两个可比较的组



“代谢综合征体检”与“长寿”是因果关系吗?

实际上，丹麦曾经进行过调查体检效果的随机对照试验。下面介绍该研究的结果。丹麦也实行与日本类似的代谢综合征体检，为国民提供糖尿病、高血压等生活习惯病的诊断及保健指导服务。随机对照试验将30～60岁的成年男女随机划分为接受体检的干预组（约1.2万人）和不接受体检的对照组（约4.8万人），用了十年时间对两组人群做了追踪调查。①

在干预组中，体检结果为未来患病风险高的人均被要求在5年内接受大约4次保健指导。结果显示，接受保健指导的大部分人在饮食习惯、运动、吸烟和饮酒习惯等方面有所改善。

然而，十年后的结果令人震惊：无论生活习惯改善与否，干预组和对照组的死亡率的差异在统计学上不具有显著性。

何谓“在统计学上具有显著性”?

“干预组和对照组的差异在统计学上不具有显著性”，是指该差异可视为巧合所导致的误差。换句话说，“观测到的差异为偶然产物的概率”低于5%，则该差异可被视为“在统计学上具有显著性”，表示两组的差异属于无法用误差或巧合解释的“具有显著性的差异”；而该概率高于5%时，则该差异“在统计学上不具有显著性”， 两组的差异可用误差或巧合来解释。（见图表2-3）5%这个数值本身并没有特殊含义，只不过长久以来经济学和统计学领域普遍采用这一水平。

图表2-3 “具有统计学意义”的小结

- 无法用误差或巧合解释的差异 → 在统计学上具有显著性
- 属于误差或巧合范围内的差异 → 在统计学上不具有显著性

我们可以再举一个更直观的例子。扔一枚硬币，正面朝上的概率是50%，反面朝上的概率也是50%；扔两次硬币，两次均为正面朝上的概率是0.5×0.5=0.25（25%）。现实中，即使两次都是正面朝上，很多人仍会认为这纯属巧合。

那么，如果扔四次硬币，还都是正面朝上，人们会怎么想？或许会开始怀疑“扔硬币的人在作弊”。那么，当第五次仍然是正面朝上时，所有人应该都会确信“这不是巧合”。

四次均正面朝上的概率是0.5×0.5×0.5×0.5=0.0625（约6%），五次均正面朝上的概率是0.5×0.5×0.5×0.5×0.5=0.03125（约3%）。判断是否在统计学上具有显著性的基准数值5%恰好处在这两个概率之间。看到投五次硬币均正面朝上，很多人会怀疑这不是单纯的巧合，而是作了弊。把这种感觉落实为实际的数值，便是5%。

换句话说，“在统计学上具有显著性”表示，两组之间的差异纯属巧合的概率就像连投五次硬币均正面朝上一样，概率非常小。

定期接受体检并不能带来长寿

丹麦的这项研究显示，接受体检未必就能长寿。不过，这项随机对照试验终归只是在丹麦国内进行的研究。评估体检是否有效是关乎国家整体健康政策的重要问题，仅参考一项随机对照试验的结果做出判断的话，会有很大的风险。因此，下面介绍几项其他国家或地区的

随机对照试验的结论。考察多项研究时，可以采用“元分析”（meta-analysis）方法。元分析是指整合多项研究结果，从整体的角度验证关系的方法（参考正文第40页）。尤其是整合多个随机对照试验的元分析在证据等级中所处等级最高。

元分析的结果已经证实，体检和长寿之间的确不存在因果关系。②

在实施随机对照试验的丹麦研究者们看来，组织大规模随机对照试验虽然成本很高，但与为全体国民提供没有效果的体检服务相比，则要划算得多。也就是说，不能贸然实施不知是否有效的政策，即使需要付出一些成本，也应该先验证因果效应是否存在，然后再决定是否要全面实施。

投入1 200亿日元税金的代谢综合征体检

尽管国外已有可靠证据表明体检不能带来长寿，日本仍在2008年启动了特定健康体检（俗称代谢综合征体检）和特定保健指导服务。

众所周知，这项政策的目的在于尽早发现并治疗生活习惯病，所有40岁以上的健康保险参保者都必须接受该体检。从2008年至2014年，日本为这项体检总计投入了大约1 200亿日元税金。

斥巨资引入的代谢综合征体检是否见效？为调查体检的效果，厚生劳动省又耗费约28亿日元架构数据库。然而数据库后来出现问题，只能分析收集到的数据中的20%，由此引发了舆论谴责。

与在全国范围内实施代谢综合征体检相比，日本完全可以先在部分地区组织随机对照试验，证实确有成效后再引入到其余地区。不同规模的随机对照试验所需成本不同，但最多也不过相当于1 200亿日元中的很少一部分。这样的话，或许国民的税金便可以用在更有效的地方。

日本既没有参考国外的先行研究成果，也没有根据本国数据验证代谢综合征体检的效果，我们对此只能表示遗憾。

这里有一点需要注意："体检"不同于"筛查"。体检为健康诊断，筛查则是针对特定疾病进行排查，如"癌症筛查"。与体检不同的是，很多筛查被证实对延长寿命具有因果效应。有证据显示，对乳腺癌、大肠癌、宫颈癌等癌症的筛查能够提高存活率。关于各类癌症的证据详情可以参考国立癌症研究中心的"推动有科学依据的癌症筛查"等页面③。

"医疗费用自付比例"和"健康"是因果关系吗？

近些年来，医疗费用增加成为日本社会普遍关注的问题。2015年，日本成为世界第三大高医疗费用支出国（占GDP比例），很多人看到这条新闻或许都会感到震惊。医疗费用高涨带来的危机感与日俱增。因此，是否应该提高老年人医疗费用的自付比例成了公众热议的话题。

2017年2月，70岁以下人口的医疗费用自付比例为30%，70～74岁人口的自付比例为20%，75岁及以上人口的自付比例为10%（各年龄段的准现役收入者④自付比例均为30%），自付比例随年龄增长而递减。与老年人自付比例为零的1973—1983年相比，各年龄段人群之间的差距已经有所缩小，但年轻群体仍抱有强烈的不公平感。

提高自付比例有望缓解稍有不适就挤占急诊资源就医的"急诊不急"现象，削减医疗费用支出。

然而，一旦提高自付比例，老年人为节省支出或许就会放弃就医。这样的话，恐怕会错过及早发现病情接受治疗的机会，给老年人的健康状况造成不良影响。

如果对老年人的健康状况没有不良影响，通过提高自付比例控制医疗费用便是合理的设想。不过，如果会带来不良影响，则需要慎重研究。在实际中，提高自付比例是否会产生不良影响呢?

一项随机对照试验——“兰德健康保险实验”（RAND Health Insurance Experiment）给出了这个问题的答案。哈佛大学医疗经济学家约瑟夫·纽豪斯在就职于美国著名智囊机构兰德公司期间实施了这项研究。该研究规模庞大，从1971年持续到1986年，以居住在美国6座城市的2 750个家庭为研究对象，投入的研究费用按现在物价计算相当于3亿美元。

组织方专门为这项研究成立了民间医疗保险公司，研究对象可免费加入医疗保险。不过，他们被随机分入四个研究组，每组的自付比例各不相同。

对照组

研究组① 自付比例0%的方案

干预组

研究组② 自付比例25%的方案

研究组③ 自付比例50%的方案

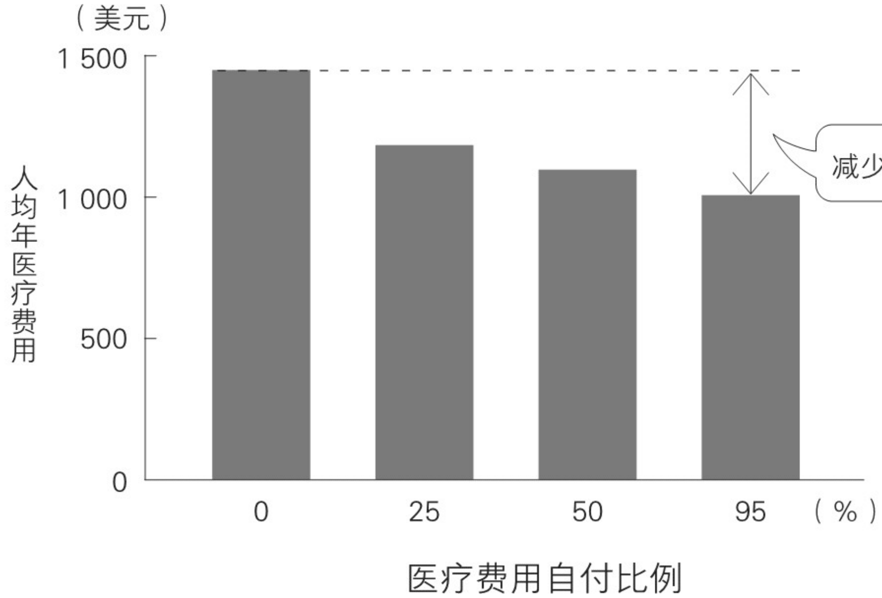
研究组④ 自付比例95%的方案

该研究的干预组多达3组，并分别设置了不同的自付比例。分入这三组的研究对象加入研究组②～④的方案，而对照组则是自付比例0%的研究组①。

兰德健康保险实验的结果

下面来看兰德健康保险实验的结果。首先是各组支出的医疗费用。如图表2-4所示，研究组①（自付比例0%）的医疗费用明显高于其他三组。和研究组④（自付比例95%）相比，高出约为30%。由此可以认为，医疗费用自付比例越高，国家整体支付的医疗费用越少。

图表2-4　医疗费用自付比例降低，医疗费用会随之增加——各种医疗费用自付比例方案的人均年医疗费用情况

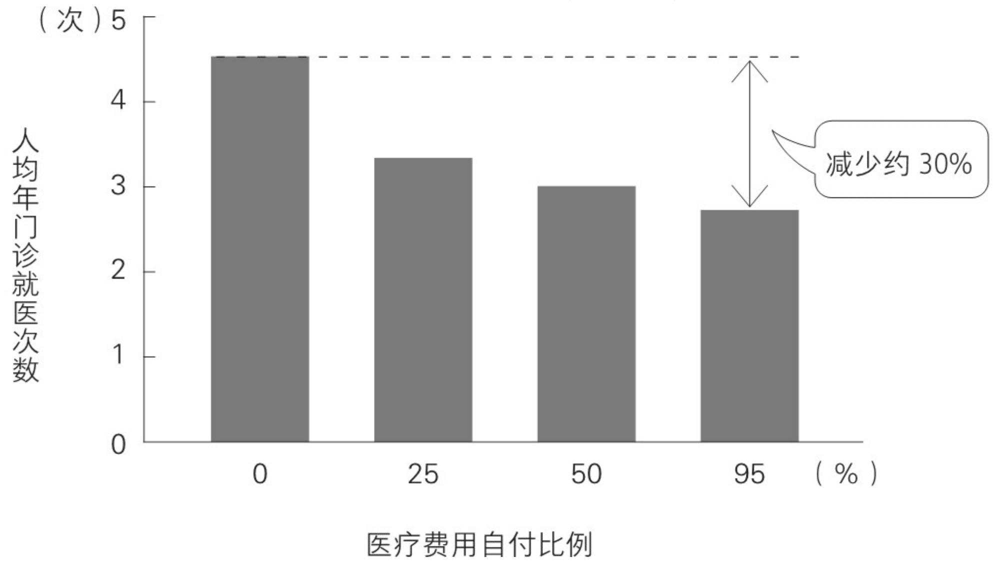


（数据来源）约瑟夫·纽豪斯等（1993）

接下来看门诊就医次数发生了什么变化。如图表2-5所示，研究组①（自付比例0%）和④（自付比例95%）之间同样存在30%左右的差异。

也就是说，医疗费用自付比例越高，人们就医或住院的次数越少。

图表2-5　医疗费用自付比例降低，门诊就医次数随之增加——各种医疗费用自付比例的人均门诊就医次数



（数据来源）约瑟夫·纽豪斯等（1993）

除贫困阶层以外，提高自付比例对健康状况没有影响

上述结果并不值得惊奇。不过，除了这些结论，兰德健康保险实验还证实了医疗费用自付比例与人们的健康状况之间不存在因果关系。在高血压症等30项健康指标上，研究组①（自付比例0%）与研究组②～④之间未发现在统计学上具有显著性的差异。

换句话说，这项研究表明，即便医疗费用自付比例提高，也不会导致人们的健康状况恶化，反而能缓解“急诊不急”现象，缩减国家整体的医疗费用支出。

不过，有一点需要注意：将研究对象限定为健康状况较差的低收入人群时，自付比例的提高就会导致人们健康状况恶化⑤。

总而言之，提高自付比例从整体来看不会对健康状况造成不良影响，但会使贫困阶层人群的健康状况恶化。在探讨提高医疗费用自付比例时，为了避免使贫困阶层受到不良影响，需要采取充分的安全保障措施，将经济困难人群的自付比例等控制在较低水平。

第2章的关键词

随机对照试验

随机对照试验通过投硬币、随机数表、抽签等方式，将研究对象人群随机分为接受干预的组（干预组）和不接受干预的组（对照组），使两组成为可比较的组，用对照组的数据替换干预组“如果没有受到干预”的反事实。

COLUMN 2
整合多项研究的“元分析”

论证某两个变量之间是否存在因果关系的研究未必只有一例。不同研究者围绕同一主题撰写论文的现象非常普遍。有些使用的数据相同，有些使用的数据不同。

如果围绕同一主题的所有研究均得出相同的结论，那倒无妨。但有时某篇论文得出的结论是“存在因果关系”，而另一篇的结论是“不存在因果关系”。遇到这种情况，不能认为只有对自己研究有利的论文结论是正确的。英语将类似的选择性失明行为称作“摘樱桃”（cherry picking），在研究中尤其需要避免。遗憾的

是，当今日本互联网上的信息汇总网站上随处可见这类“摘樱桃”的行为，错误信息广为扩散的现象也不在少数。

这时需要用到的就是“元分析”（meta-analysis）。“meta”意为“高层次的”，“ analysis”意为“分析”，是一种综合多项研究结果，从整体上论证变量间关系的研究方法。

近年来，日本国立癌症研究中心和日本烟草公司（JT）的对峙等使元分析得到了广泛关注。

国外已有研究证实被动吸烟确实会增加肺癌的患病风险。为此，欧美的很多国家制定了法律，全面禁止在公共场所、餐厅等室内空间吸烟。然而，用日本人的数据进行的研究，尚未得出能够证明被动吸烟与肺癌之间存在因果关系的结论。虽然已经出现了一些以日本人为对象的研究，但由于对象人数较少，未能得出在统计学上具有显著性的结果。

于是，日本国立癌症研究中心的研究团队于2016年8月发表了一份整合了9项采用日本国内数据的观察研究的元分析。该分析的结论是：对日本人来说，被动吸烟也会使肺癌患病风险提高1.3倍。据此，日本国立癌症研究中心指出，鉴于已明确证实被动吸烟会导致不吸烟的日本人肺癌患病风险增高，有必要全面禁止室内吸烟行为，仿效国外采取预防被动吸烟的措施。

然而，日本烟草公司作为日本烟草行业的代表对此结论提出了质疑。在日本国立癌症研究中心发表元分析结果的当天，日本烟草公司便以社长名义发表抗议声明，称该元分析“把时期和条件各不相同，且结果均在统计学上不具有显著性的9项研究整合在一起”，其结果“难以证明被动吸烟与肺癌之间确实存在因果关系”。

对此，日本国立癌症研究中心的研究者们迅速反驳，批判日本烟草公司的声明“轻视了被动吸烟的危害”，称他们并非为了利于得出该结论才选择了这9项研究，而是从采用日本人数据的论文中选出所有提及因果关系的论文，并依照科学步骤对其进行综合分析。研究中心用确凿的证据断然反驳了日本烟草公司的主张，

指出被动吸烟并非日本烟草公司所称的“添麻烦”或者“素质”层面的问题，而是“拥有科学依据的、危害健康的问题”。最终，被动吸烟会提高肺癌患病风险的认知由此得到了普及。

① 在这项研究中，对照组的人数是干预组的4倍，不过这里的重点是分组决定方式的随机性，两组人数不需要完全相同。

② 看到这个结论便认为体检没有任何意义还为时过早。虽然体检不能带来长寿，但在糖尿病或高血压的初期阶段接受治疗可预防失明、脑梗死等并发症，从结果来看，可以提高生活质量。

③ 国立癌症研究中心的主页（http://canscreen.ncc.go.jp/）。

④ “准现役收入者”指按照日本医疗制度规定计算的年应税收入超过145万日元的70岁以上老年人。

⑤ 结果显示，30项健康指标当中，与高血压症、视力、牙科护理、危重症状等4项相关的健康状况出现恶化。“危重症状”指胸痛、出血、失去意识、呼吸困难、体重减少10磅（约4.5千克）以上等症状。

第3章　男医生比女医生更优秀吗?
利用与实验类似的偶发现象进行自然实验

用现有数据重现与实验类似的环境

随机对照试验的证据等级很高，但实施过程中会遇到不少困难。比如，可能很难对作为研究对象的个体或企业进行人为随机分组。研究对象可能会强烈反对，或者很多情况下出于伦理考虑，无法进行随机对照实验。

让我们用珠宝店的例子设想一下。你所在的珠宝店在全国共有100家店铺。如前文所述，你想明确广告给销售额带来的因果效应。如果可以组织随机对照试验，只需把100家店铺随机分成投放广告的店铺（干预组）和不投放广告的店铺（对照组），进行比较即可。然而实际上这恐怕不会这么轻易实现。

被分入不投放广告组的店铺可能会有意见，在重视速度的销售最前线，恐怕也没有充足的时间等待随机对照试验的结果出来之后再做决定。

再者，如今已是大数据时代。对很多情况，我们手头已经掌握了各种各样的数据。如果对这些数据视而不见，只会感叹“不做随机对照试验，就无法测定广告效果”，这样的人实在称不上有能力的企业人。

无法实施随机对照试验等人为实验时，我们只能用手头现有的数据进行分析。通过随机对照试验得到的数据称为“实验数据”，通过日常经济活动得到的数据以及从政府统计调查等渠道获得的数据称为“观测数据”。

采用观测数据进行分析时，首要工作是从观测数据中找出“与人为实验相类似”的环境。本章介绍的“自然实验”，便是通过法律或制度变化、自然灾害、战乱等人们无法预测的变化，找到与随机对照试验类似的环境，从而明确两个变量之间的因果关系。

此处重申，这种方法的目的同样也是“制造可比较的组，用最贴切的值替换反事实”。事先无法预测的某些因素（经济学称其为“外生冲击”）可以在不经意中将研究对象自然而然地划分为干预组和对照组，自然实验就是要利用这种环境。根据研究内容的不同，可以利用的外生冲击包括制度变革、自然灾害等。

那么回到上面的例子，考虑到实施随机对照试验的难度，你决定为所有的100家店铺统一投放报纸折页广告。可是，负责印刷广告的某家企业印刷机出现故障，导致部分店铺未能投放广告。

于是你想到了一个主意：利用这种情况，对如期投放广告的店铺和未能投放广告的店铺进行比较，是不是就能评估出广告给销售额带来的因果效应呢？实际上，这正是自然实验的思维方式。

“医生性别”与“患者死亡率”是因果关系吗?

下面介绍一项运用自然实验的研究。走进书店，能看到书架上摆着很多介绍“名医排行”的书籍。翻过这类书的人或许会发现，绝大多数“名医”都是男性。人们似乎大多抱有一种印象，认为名医肯定是男医生。然而，男医生真的比女医生更优秀吗?

本书的作者之一津川曾研究过这一课题。他分析了2011年至2014年因内科疾病在美国医院住院的超过100万名患者的数据，试图验证主治医生为男性的情况下和主治医生为女性的情况下，患者住院后30天之内死亡的概率（30天死亡率）有无差异。

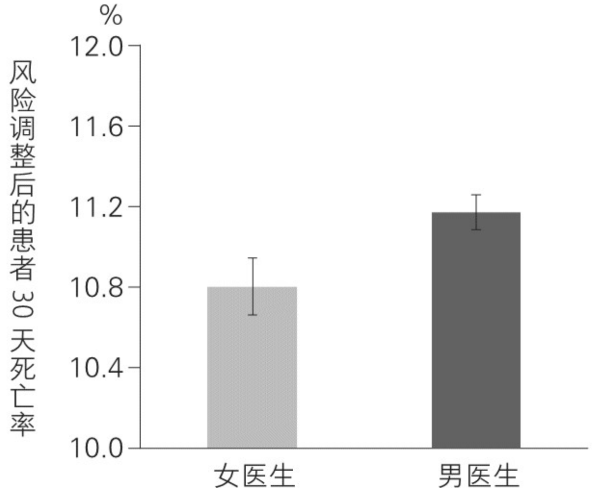
思考这个问题时必须注意一点：男医生或许更倾向于选择重症患者，或者重症患者更倾向于选择男医师做自己的主治医生。为排除这类选择性偏差的干扰，津川等人选择针对住院医师（hospitalist）进行研究。

住院医师指不对外坐诊、只负责治疗住院患者的内科医生。这类专业医生自20世纪90年代起在美国得到飞速发展，如今绝大多数大型医院的内科住院患者都由住院医师负责治疗。住院医师的工作通常为轮班制，主要负责自己当班时收治的患者。也就是说，住院医师无法选择自己负责哪些患者，而患者也不能选择自己的主治医生。这种情况正可谓将患者随机分配到男医生组或女医生组的自然实验。

女医生负责的患者死亡率更低

津川等人的分析结果显示，女医生负责的患者的30天死亡率比男医生的患者低0.4%。（见图表3-1）

图表3-1　女医生负责的患者30 天死亡率更低



（注）本数据为对患者重症程度、医生特点等因素进行调整后，对就职于同一所医院的男医生和女医生（均为住院医师）进行比较的结果。柱状图顶部的竖线为“95% 置信区间”，表示推测值有95% 的概率属于该区间范围内。更严密地说，95% 置信区间的定义是：将同一研究重复进行100 次，每次均构造出95% 置信区间，其中有95 次的置信区间会包含真值。

（数据来源）摘自津川等（2017），对部分内容稍做了调整。

也许有人觉得0.4%的差异很小，几乎可以忽略不计。但其实这个差异绝对不小。死亡率0.4%的差异和近十年美国通过各种努力才实现的住院患者死亡率的下降幅度几乎持平。美国通过研发新药和医疗器械、开展医学研究获得最新知识、完善临床指南等途径实现的死亡率降幅和男女医生负责患者的死亡率之差相当，这个事实足以让人震惊。

为什么女医生负责的患者死亡率更低？既往研究发现，女医生遵照临床指南进行治疗的比例更高，她们和患者之间的交流也更为密切。有可能是男女医生在治疗方法上的差异导致了患者预后的差异。①

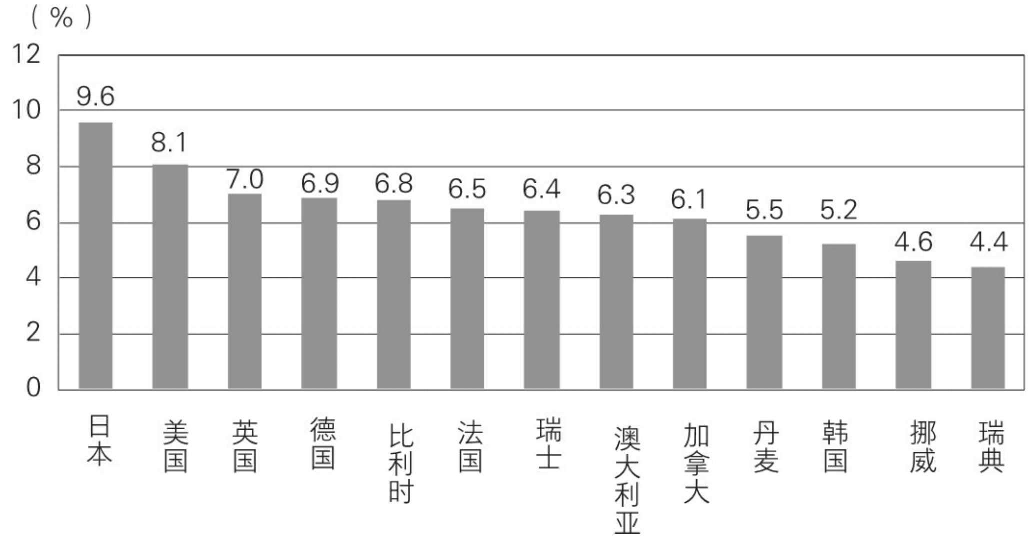
在美国，与男医生相比，女医生薪酬更低，晋升所需年数更多，这已经上升为社会问题。从该研究中也可以发现，女医生提供的治疗质量高于男医生。因此，在医疗领域，也需要消除性别带来的差距。

“出生体重”与“健康”是因果关系吗？

日本有这样一句老话：先生小娃再养大。说的是婴儿出生时体重越轻越好，生下后再把孩子养胖更好。在剖宫产技术尚不成熟的年代，为降低产妇因为生产而死亡的风险，才会出现这句老话。

或许是受这种观念的影响，日本的低出生体重儿（出生体重低于2 500克）比例要高于其他国家。（见图表3-2）

图表3-2　日本的低出生体重儿多于他国——各国低出生体重儿的比例



（注）低出生体重儿的定义为出生体重低于2 500 克的婴儿（2011 年）。

（数据来源）OECD Health Statistics 2016

低体重对婴儿真的没有影响吗？一项经济学研究运用双胞胎的数据探讨了这个问题，其方法是对双胞胎中出生体重较重的婴儿（干预组）和出生体重较轻的婴儿（对照组）进行比较。

也许有人会想双胞胎的体重也存在差异吗？一般而言在母体中营养状况相对更好的胎儿体重会更重，也会先出生。当然，双胞胎的出生顺序是偶然的，他们自然分为体重较重的婴儿和体重较轻的婴儿也可看作一种自然实验。

出生体重较重的婴儿更健康

美国、挪威、加拿大、中国台湾采用大量双胞胎数据展开的研究证实，出生体重较重的婴儿长大后的成绩、学历、收入和健康状况更好。本书作者之一的中室曾利用日本的双胞胎数据进行研究，也发现出生体重较重的孩子初中毕业时的成绩更好。从孩子的角度来考虑，“先生小娃再养大”的建议并不对。

近年来，不少经济学研究成果支持“胎儿起源说”。胎儿起源说认为，胎儿时期的发育环境对之后的人生具有决定性的重要影响。这一主张和出生体重与成绩、学历、收入及健康状况之间存在因果关系的经济学研究成果并不矛盾。它不仅阐述了确保孕产妇过上多彩且舒适生活的重要性，也为探讨社会应如何保护和支援准妈妈提供了契机。

基于“胎儿起源说”撰写了畅销书《胎内人生》的作家安妮·墨菲·保罗（Annie Murphy Paul）在TED演讲中曾说过这样一番话：“研究胎儿起源不是为了追究准妈妈在怀孕过程经历的不幸遭遇或行为，而是为了帮助下一代的孩子们发现更美好的人生。”

第3章的关键词

自然实验

利用研究对象人群由于法律制度变更、自然灾害等“外生冲击”的影响而自然分成受影响组（干预组）和不受影响组（对照组）的现象，来验证因果关系。

COLUMN 3
被动吸烟会增加心脏病的患病风险吗?

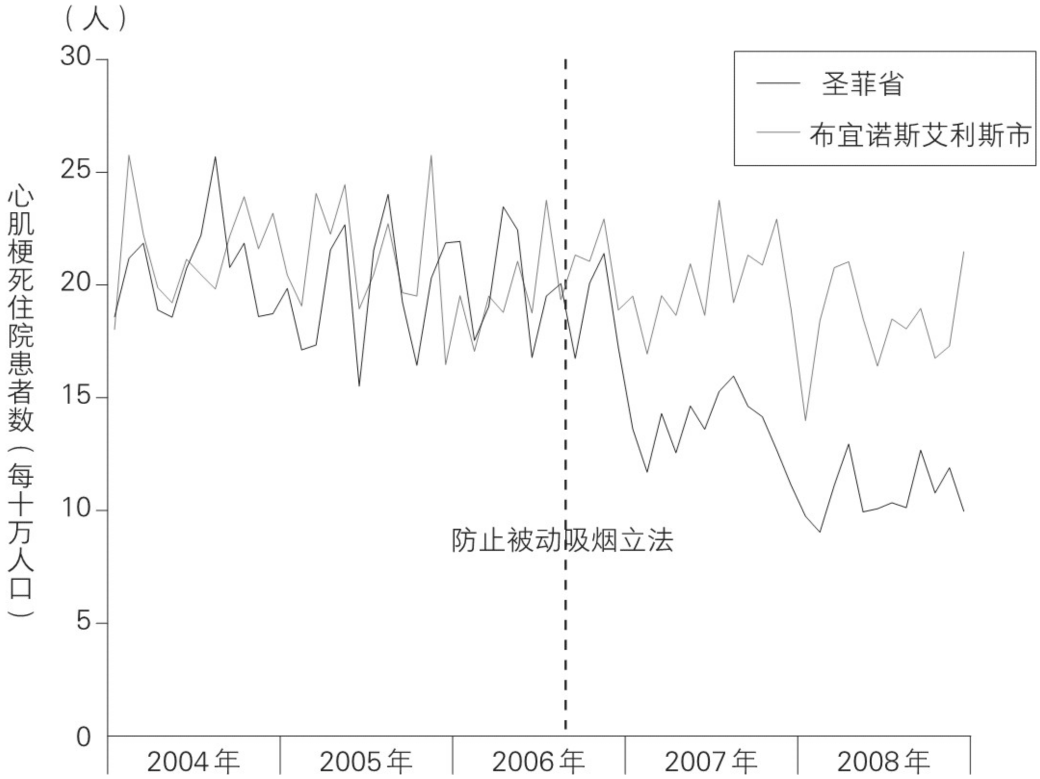
日本国立癌症研究中心实施的元分析研究证实了被动吸烟与患肺癌之间确实存在因果关系。不过被动吸烟引发的疾病不只有肺癌。使用阿根廷的数据进行的一项研究表明，被动吸烟与心肌梗死之间也存在因果关系。

自2005年世界卫生组织（WHO）批准《烟草控制框架公约》以来，阿根廷迅速加强了对烟草的控制。然而，随着阿根廷地方分权制度的深化，各省在政治和财政方面掌握的自治权大于国家，烟草控制的落实情况也各不相同。

例如，位于阿根廷北部的圣菲省颁布了严格的禁烟法令，规定自2006年8月起公共场所全面禁止吸烟，甚至还设立了举报热线，吸烟者、知情却未加以阻止的餐厅或酒吧都属于举报对象。其惩罚措施非常严厉，一旦违反该规定，不仅吸烟者本人，就连默许其吸烟的店铺也必须交纳罚金，甚至被迫停业。而相比之下，阿根廷首都布宜诺斯艾利斯市的规定则比较宽松，餐厅或酒吧只要安装换气设施，便允许吸烟。

阿根廷保健部的研究人员试图把这种情况视为自然实验，对控制严格的圣菲省（干预组）和控制宽松的布宜诺斯艾利斯市（对照组）进行比较。该项研究得出的结果引人深思：两个地区在实施控制规定后，吸烟率并未发生变化。即实行控制未能阻止吸烟者继续吸烟。不过，实行严格控制的圣菲省的心肌梗死住院患者人数要比布宜诺斯艾利斯市少13%（图表3-3）。也就是说，吸烟者未发生变化，但被迫受二手烟影响的人的健康状况得到了改善。

图表3-3　公共场所禁烟有利于非吸烟者的健康——圣菲省和布宜诺斯艾利斯市的心肌梗死住院患者数



（数据来源）本图表由作者根据费兰特（Ferrante）等人的研究（2012）制成。

不过，严格控制吸烟没有对圣菲省的经济造成不良影响吗？规定实行初期，圣菲省的餐厅及酒吧经营者曾担忧吸烟者会减少外出就餐，使营业额下降。然而，后来有研究比较了规定严格的省份和规定宽松的省份餐厅及酒吧的营业情况，发现两个地区的营业额差异在统计学上不具有显著性。美国也有很多州和城市实行了相同规定，过去的研究已经证明，这些规定不仅不会影响餐厅或酒吧的营业额，对宾馆等行业的销售情况也没有影响。

众所周知，日本的被动吸烟防范制度及措施尚不到位，从这项研究可知，布宜诺斯艾利斯市的部分被动吸烟防范措施不足以保护不吸烟者免受二手烟的伤害。日本也应该制定更为严格的规定防止被动吸烟，这个建议绝不过分。

① 这项研究的对象为内科医生。目前尚无研究表明研究外科或其他领域的医生会得出相同结果。

第4章　最低工资与就业之间存在因果关系吗?排除趋势影响的双重差分法

模仿实验的“准实验”

自然实验是利用观测数据和外生冲击，找到类似随机对照试验的环境。然而，日常生活中很少出现法律制度变更、自然灾害等外生冲击，很难找到可视为自然实验的环境。在这种情况下，我们应该怎么做呢?

本章开始介绍“准实验”。准实验是对实验的“模仿”，即通过观测数据和统计学方法，打造出与实施随机对照试验相类似的状态。关于这里所说的“统计学方法”，本书将介绍以下4种。

- “双重差分法（第4章）”
- “工具变量法（第5章）”
- “断点回归法（第6章）”
- “匹配法（第7章）”

前后比较毫无意义

让我们回到珠宝店的例子。你所在的珠宝店在2014年圣诞节期间没有投放广告，当年12月的销售额为1 000万日元；而2015年的圣诞季商战中大规模投放了报纸广告，当年12月的销售额是1 400万日元。

减去投放广告所花费的100万日元，2015年12月的销售额比上一年同期增长了300万日元。（见图表4-1）

图表4-1　通过前后比较，能推算出广告的效果吗?

	2014 年 12 月	2015 年 12 月	差异
销售额	1 000 万日元	1 400 万日元	400 万日元
有无投放广告	×	○	
广告费用	0 日元	100 万日元	100 万日元

销售额增长 300 万日元！

我们可以把这300万日元归功于广告吗？2016年的圣诞节之前应该继续投放广告吗?

倘若只参考图表4-1，我们很可能会得出这样的结论：销售额从2014年的1 000万日元增长到2015年的1 400万日元，这是广告效应，2016年应当继续投放广告。这种单纯比较投放广告前后结果的分析方法叫作“实验前后测设计”。可能有人常会看到这种分析方法被用来论证某些措施或策划确实产生了功效。

然而，实验前后测设计无法证明广告与销售额之间存在因果关系。

不能使用实验前后测分析的两个原因

为什么不能使用实验前后测分析？原因有二。第一，这种方法没能考虑到随时间产生的自然变化（“趋势”）所造成的影响。以珠宝店为例，2015年销售额增长也许只是因为当年经济形势更好，即使不投放广告或许也能卖出1 400万日元。“趋势”的形成与投放广告与否无关，但实验前后测分析容易将其误视为广告的效果。

第二，是“回归平均”的可能性。这是一种统计学现象，在不断收集数据的过程中，偶尔出现一个极端值后，数据会逐渐回归到通常的水平。也许有人遇到过这样的事：偶尔测量一次血压，可能会出现很高或很低的值，但只要再多测几次，就会变成平时的血压。当然也可能是跑完步等情况下血压变高，不过在反复测量的过程中“回归平均”（平时血压值）。说不定你所在的珠宝店也只是碰巧在2014年销售额下降。如果真是这样的话，销售额之后便会回归平均，因此2015年销售额增加也不足为奇。

投放广告后，即使销售额有所增长，我们也无法断定这不是“趋势”或回归平均现象所导致的结果。因此，实验前后测分析无法用来评估广告的因果效应。

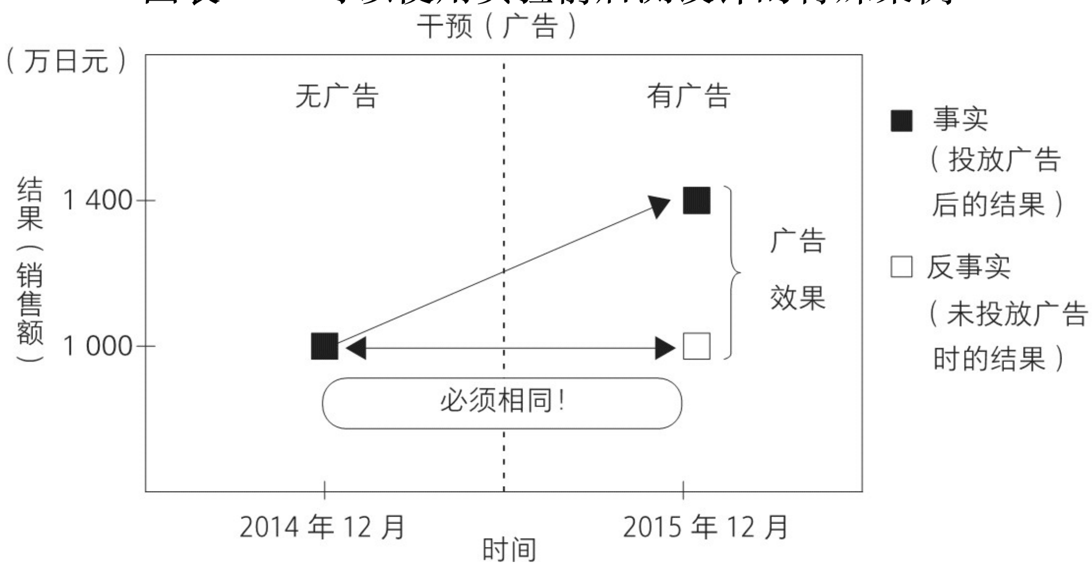
去年的销售额为“反事实”时，前后比较才有效

不过，实验前后测设计在一些特殊案例中也可以使用。我们继续思考图表4-1的事例。

要证明2014年和2015年的销售额之差（400万日元）是广告的因果效应，必须满足一个条件，即2015年如果没有投放广告，就会实现和2014年完全相同的销售额。

换句话说，2014年的销售额必须可以代表“如果2015年未投放广告”的反事实。（见图表4-2）

图表4-2　可以使用实验前后测设计的特殊案例



然而，实际上无论投放广告与否，销售额都会受到“趋势”的影响或者增加，或者减少。图表4-3是从图表4-1追溯到过去几年的数据。从图表中可以看出，2012年至2015年，你所在的珠宝店很受欢迎，销售额连年攀升。尽管2012年到2014年从未打过广告，销售额仍以每年400万日元的增速不断上涨。

图表4-3　往年销售额数据回顾

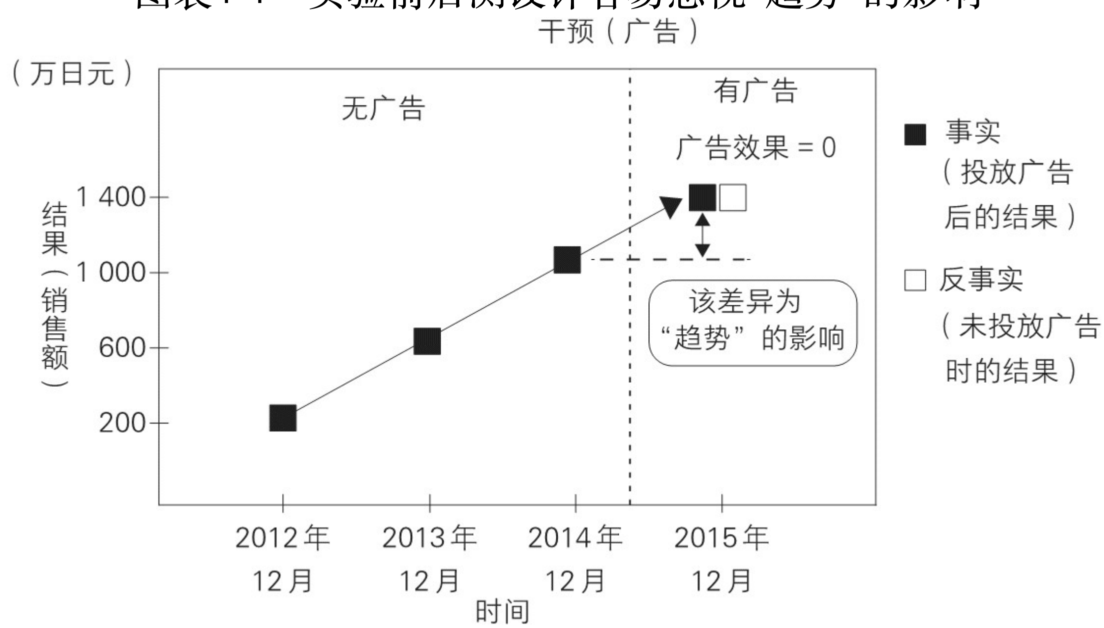
时间	2012 年 12 月	2013 年 12 月	2014 年 12 月	2015 年 12 月
销售额	200 万日元	600 万日元	1 000 万日元	1 400 万日元
有无广告	×	×	×	○

+400 万日元　+400 万日元　+400 万日元

这种情况下，2015年较2014年增长的400万日元销售额就不是广告的效果，而只是“趋势”。如果没有看到2012年到2014年的销售额走势，恐怕就会把2014年到2015年增长的400万日元误认为是广告的效果了。

再来看图表4-4。该图表体现了无论在事实还是在反事实中，销售额都不断增长的“趋势”，说明广告效果为零。

图表4-4　实验前后测设计容易忽视“趋势”的影响



实验前后测设计的改良版——双重差分法

“双重差分法”是实验前后测设计的改良版。它与实验前后测设计不同之处在于，双重差分法需要代表反事实的“对照组”。

要采用双重差分法，必须分别获取干预组与对照组在实验前后两个时点的数据。顾名思义，这种方法需要两个“差异”数据。一个是干预前后的“差异”（与实验前后测设计推算的效果相同）；另一个则是干预组与对照组的“差异”。这种方法用两个“差异”的差异来推算干预的效果，故名“双重差分法”。

下面进一步介绍这种方法。如图表4-5所示，在受到干预的组（干预组）中，设干预前的结果为A1，干预后的结果为A2；然后在未受到干预的组（对照组）中，设干预前的结果为B1，干预后的结果为B2。

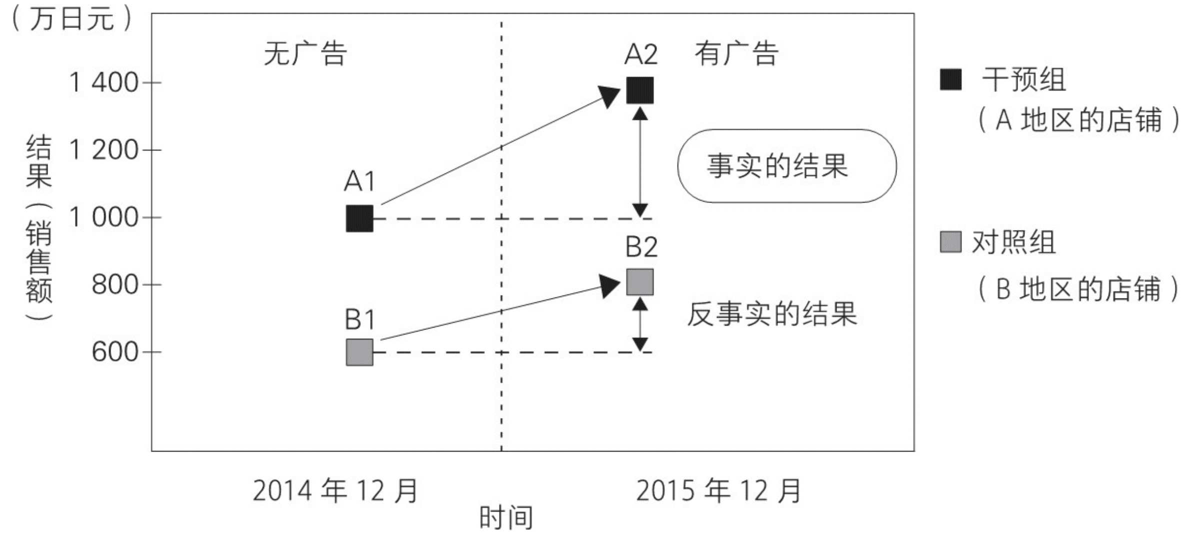
图表4-5　双重差分法的概念（1）

	干预前	干预后	差异
干预组	A1	A2	A2−A1
对照组	B1	B2	B2−B1

二者之差 {(A2−A1)−(B2−B1)} 即因果效应

干预组的前后比较（A2-A1）和对照组的前后比较（B2-B1）之间的差（A2-A1）-（B2-B1），即为用双重差分法推算出的干预效果。图表4-6中，从A1到A2的线代表事实（投放广告后的结果），从B1到B2的线代表反事实（如果投放广告的店铺未曾投放广告会产生什么结果）。

图表4-6　双重差分法的概念（2）



从干预组的前后差异A2-A1中减去对照组的前后差异B2-B1，这种方法排除了“趋势”的影响，所以能够准确地推算出因果效应。

以珠宝店为例具体说明。假设在全国所有店铺中，A地区的店铺在2015年投放了广告，而B地区的店铺在同一时期没有投放广告。B地区店铺2014年12月的销售额为600万日元，2015年12月的销售额为800万日元。（见图表4-7）

图表4-7　用双重差分法推算珠宝店的销售额

	2014 年 12 月	2015 年 12 月	差异
干预组的销售额（A 地区的店铺）	1 000 万日元	1 400 万日元	400 万日元
对照组的销售额（B 地区的店铺）	600 万日元	800 万日元	200 万日元

200 万日元的因果效应

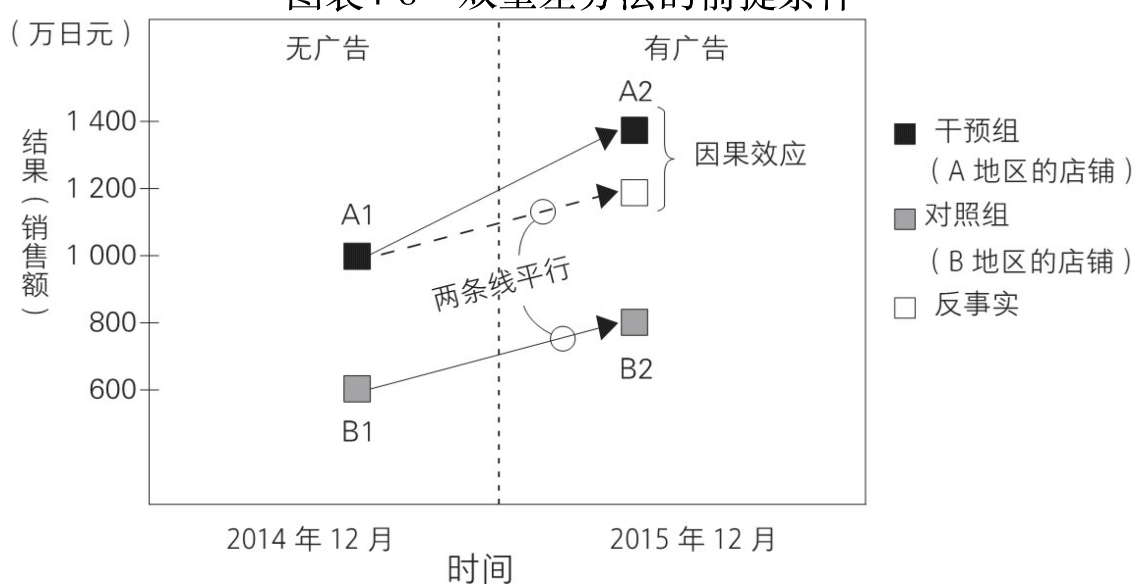
投放了广告的A地区店铺2015年的销售额相比2014年增长了400万日元，而没有投放广告的B地区店铺销售额增长了200万日元（800万日元-600万日元）。两地区的销售额涨幅的差200万日元（400万日元-200万日元），即为双重差分法得出的干预的因果效应。如前文介绍的，即使广告投放成本为100万日元，投放广告也能为店铺带来100万日元（200万日元-100万日元）的额外收益。

双重差分法成立的两个前提条件

双重差分法成立必须满足两个前提条件。

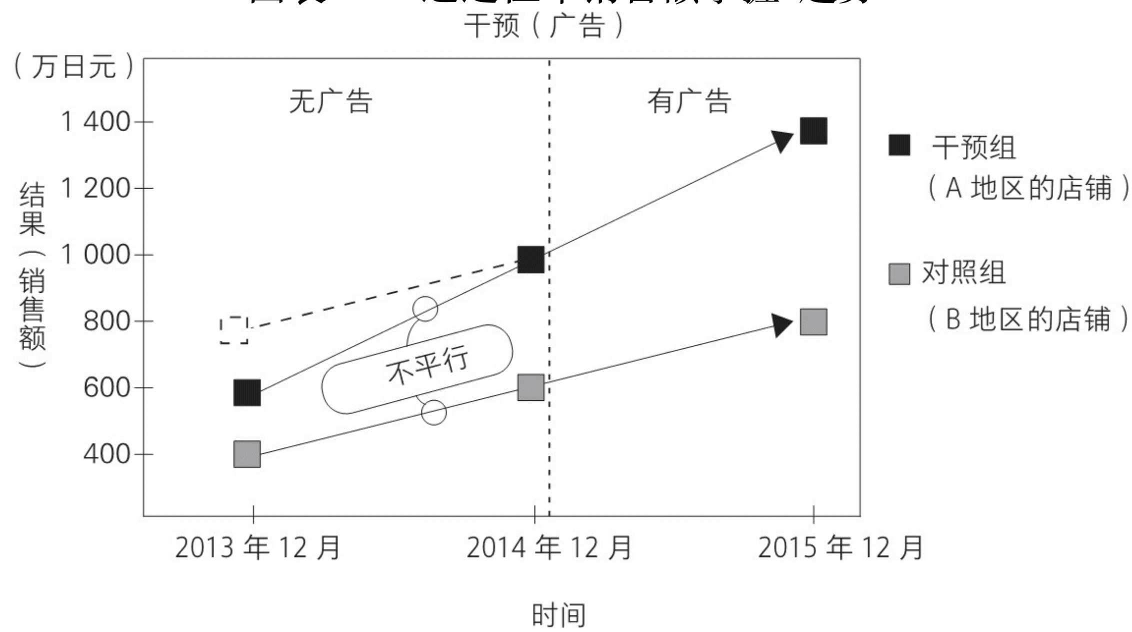
第一，“干预组和对照组在投放广告之前的销售额趋势平行”。B地区店铺的销售额代表了A地区店铺如果未投放广告时的反事实状态。因此，A地区店铺和B地区店铺至少在干预前必须为“可比较”的组。换句话说，A地区和B地区在投放广告之前的销售额必须具有相同的“趋势”（图表4-8中体现的倾斜程度）。

图表4-8　双重差分法的前提条件



不过，只看2014年和2015年的数据，无法判断A地区和B地区在投放广告之前的销售额“趋势”是否相同。于是，我们再来看一看2013年12月的数据。（见图表4-9）

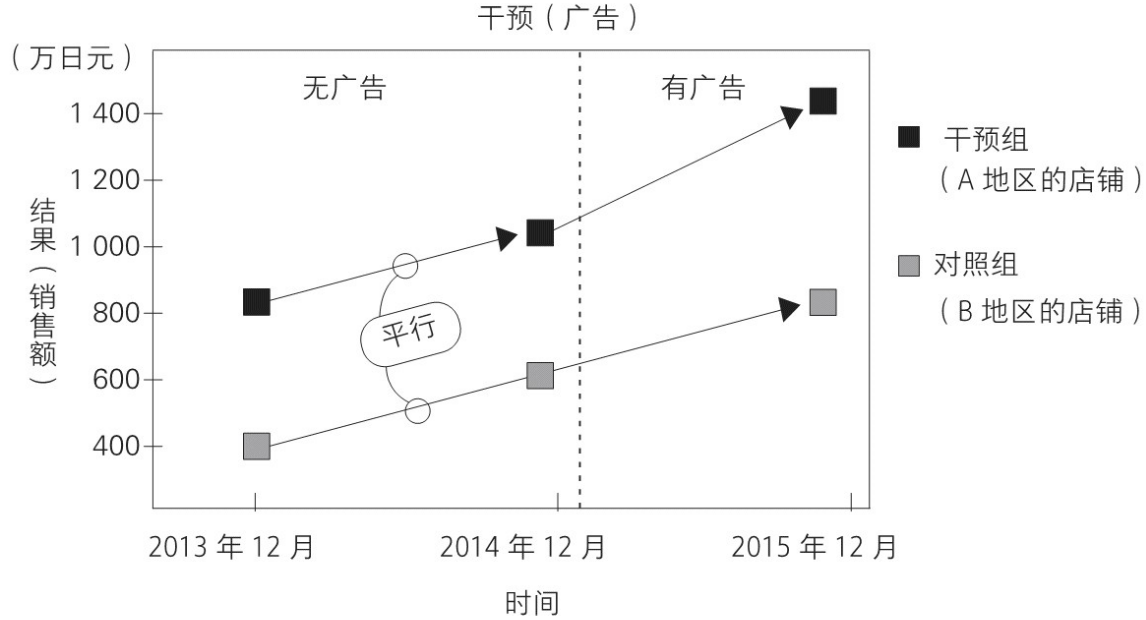
图表4-9　通过往年销售额掌握“趋势”



从图表4-9中不难看出，在投放广告之前，即2013年12月到2014年12月期间，A地区和B地区的销售额“趋势”并不相同。也就是说，A地区的店铺即使不投放广告，每年销售额也会持续增加400万日元，而B地区的店铺销售额每年涨幅只有200万日元。这样来看，两地区的店铺不满足“干预组和对照组在投放广告之前的销售额趋势平行”的前提条件，因此无法使用双重差分法。

另一方面，我们再来看一看图表4-10。A地区和B地区在投放广告前，销售额的“趋势”是相同的。二者都以每年200万日元的幅度增长。这种情况满足“干预组和对照组在投放广告之前的销售额趋势平行”的前提条件，因此可以使用双重差分法。

图表4-10 “趋势”平行，则可以使用双重差分法



第二个前提条件是，受到干预期间（在该事例中即投放广告的2014年12月到2015年12月之间）没有出现影响销售额的“其他变化”。

例如假设在2015年11月播出的电视剧中，当红女星佩戴的一款项链受到人们的关注。然而，该电视剧只在A地区播出，这款项链也只在A地区热销。这样就会有问题，使我们无法判断双重差分法推算出的200万日元销售额增幅是广告带来的效果还是电视剧带来的效果。

“保育所数量”与“母亲就业”是因果关系吗?

下面介绍一项应用双重差分法的研究。

“没上成保育园，日本去死！”这句话登上了2016年日本U-CAN新语及流行语大奖的候选名单。因孩子上不成保育园①而不得不辞去工作的母亲的这句肺腑之言使日本社会意识到保育园入园难问题的严重

性。这句话引发的关注也促使政府制定了紧急措施，放宽了针对保育园的规定。

然而，增设认可保育所真的能促进母亲就业吗？这个问题还有待慎重探讨。因为有研究发现，挪威、法国、美国等国家虽然配备了认可保育所，但母亲就业率并未提升。

对这个问题，我们也需要认真思考保育园与母亲就业的关系属于因果关系还是相关关系。究竟是“因为有保育园，所以母亲去就业”（因果关系），还是“母亲就业率高的地区保育园数量也多”（相关关系）？

增设认可保育所不会提升母亲的就业率

东京大学的朝井友纪子、一桥大学的神林龙和麦克马斯特大学的山口慎太郎针对这一问题展开了研究。他们采用1990年到2010年间日本各行政区的保育园定员率②和母亲就业率数据进行了双重差分法分析。第一个差异是1990年到2010年间日本各行政区母亲就业率差异，第二个差异是保育园定员率有所增长的行政区（干预组）和保育园定员率完全或几乎没有增长的行政区（对照组）的母亲就业率差异。朝井等人试图用这两个“差异”推算出保育园定员率的增加对母亲就业率提升的因果效应。

分析结果令人震惊：保育园定员率与母亲就业率之间未发现因果关系。究其原因，朝井等人认为认可保育所可能只起到了替代私人保育服务（祖父母、保姆或认可外保育所等）的作用。就业意愿较强的女性原本就在利用这些私人保育服务来确保自己继续工作。因此，认可保育所定员的增加只是促使这些女性由利用私人保育服务改为利用公共保育服务，而并未起到促进全职主妇就业的作用。

最与保育园形成替代关系的恐怕要属祖父母了。据日本厚生劳动省2003年的统计，3岁儿童“平时的保育者”（可多选）中，保育园的保育师占27%，而祖父母占比则达到38%，女方父母的比例尤其较高。

这篇论文的作者之一山口指出，即使增设认可保育所并不会对母亲就业产生因果效应，但保育园是由拥有专业知识和技能的保育员所组成的专业组织，可能对儿童的成长和健康产生更多的积极影响。实际上，以往很多经济学研究已证实，优质幼儿教育的投资回报率极高。从这个角度来看，或许正如山口所言，增设认可保育所的目的与其说是促进母亲就业，不如定位为为孩子开拓更好的未来更恰当。

“最低工资”与“就业”是因果关系吗?

再来看一项应用双重差分法的研究。

雇主必须向劳动者支付报酬的最低限额被称作“最低工资”。日本各地的最低工资标准不尽相同。例如，2016年10月，东京都的最低工资为每小时932日元，冲绳县的最低工资为每小时714日元。安倍内阁的成长战略也提及由政府主导提高最低工资。对此除了欢迎的声音，也有不少人担忧它会对就业造成不良影响。

如果企业决定通过裁员调整人员结构，以此抵销提高最低工资所导致的成本增加，那么最低工资的提高确实会导致就业机会的减少。实际上，据说20世纪70年代的美国就曾经出现过这种情况。

不过，断定最低工资与就业之间存在因果关系还为之过早。提高最低工资往往是经济形势恶化时所采取的政策，其目的在于提高工资，带动个人消费。这种情况下，“经济形势恶化”是同时影响最低工资和就业的混杂因素，使我们无法判断到底是“提高最低工资导致了雇佣减少”（因果关系），还是“就业环境恶化导致就业机会减少，于是提高最低工资”（相关关系）。

向这个问题发起挑战的是加州大学伯克利分校的大卫·卡德（David Card）和普林斯顿大学的阿兰·克鲁格（Alan Krueger）。二人关注了新泽西州和宾夕法尼亚州交界处相邻的两个县。在美国，最低工资标准的调整以州为单位各自执行，1992年新泽西州将最低工资从每小时4.25美元提高至5.05美元，而宾夕法尼亚州则保持原有标准不变。

这项研究中的第一个差异是1992年前后两个州的就业率差异，第二个差异是新泽西州（干预组）和宾夕法尼亚州（对照组）的就业率差异。二人试图通过这两个“差异”推算出提高最低工资对就业的因果效应。

提高最低工资不会减少就业

卡德和克鲁格的分析发现，提高最低工资不会减少就业。[3]此外还发现，提高最低工资导致了新泽西州企业提高了产品价格。也就是说，企业并未通过裁员来克服提高最低工资带来的成本上涨，而是将其转移到产品价格上。

马萨诸塞大学安姆斯特分校的杜布（Dube）等人在论文中把新泽西州和宾夕法尼亚州的研究扩大到全美，同样未能发现最低工资会对就业产生因果效应。他们认为最低工资的平稳上升对就业的不良影响是有限的。

第4章的关键词

双重差分法

该方法从受到干预的组（干预组）和未受干预的组（对照组）中获取干预前后的结果差异，以及干预后和对照组的结果差异。不过，该方法只在两个前提条件成立时才能发挥作用。第一个前提条件是，干预组和对照组在干预前的结果具有相同“趋势”，即“趋势”是“可比较”的。第二个前提条件是，在施加干预期间，没有其他变化对干预组和对照组的结果产生影响。

COLUMN 4
“不快点睡觉，妖怪就要来了”是正确的教育方法吗?

美国有一种名为“恐吓从善”（scared straight）的教育方法。通过恐吓让孩子认识到遵守规矩的必要性。例如通过重现交通事故现场让孩子学习遵守交通法规的重要性，或者带不良少年参观监狱以促使他们改过自新。

日本也有一些父母会为了督促孩子睡觉，跟孩子说“不快点睡觉，妖怪就要来了”。看来“恐吓从善”并不是美国人特有的教育方法。

不过，美国的“恐吓从善”之所以更为出名，是因为20世纪70年代，电视曾经报道过一个青年团伙接受这种教育后不再犯罪的事例。从那以后，便有很多人形成了这样的认知：“恐吓从善”具有抑制年轻人犯罪的功效。

这是仅仅通过前后比较，用实验前后测设计（参考正文第58页）推导出错误结论的典型事例。电视报道的情况或许只是年轻人长大后认识到自己过去的所作所为非常愚蠢可笑，或者是电视

台为了制作节目专门找了一些行为尤其恶劣的年轻人，而后来则出现了“回归平均”（参考正文第59页）。

后来，研究人员对接受过“恐吓从善”教育的年轻人（干预组）和未接受该教育的年轻人（对照组）进行了随机分组比较试验，发现了一个惊人的事实：与未接受该教育的年轻人（对照组）相比，接受过该教育的年轻人（干预组）在后来的人生中涉足犯罪的概率更高。

也就是说，该教育项目看似有效，实际上不仅不具备让青年人改过自新的力量，反而提高了他们犯罪的概率。

采用实验前后测设计，轻率地评估政策效果的做法和“恐吓从善”教育的案例一样，不仅得不到预期的结果，搞不好还会对有可能成为社会危害的政策做出好评。这样的政策在日本比比皆是，例如“宽松教育”，一开始便在全国范围内实行，然后像过时的流行一样不了了之，之后只采用实验前后测设计进行政策评估。不验证因果关系，便推行看似有效的政策，无异于将巨大的风险转嫁到国民身上，这一点我们可不要忘了。

① “保育园”为各种“保育所”的统称，指根据日本的《儿童福祉法》，以“为缺乏看护的婴幼儿提供保育服务”为目的，经有关机构批准设立的机构，一般以0岁至小学入学前儿童为保育对象。本书中的“认可保育所”为其中符合“保育所最低标准”，经过日本各地方行政机构的知事认可的保育园。而后文的“认可外保育所”则指未经过地方行政机构认可的保育园。

② 保育园定员率为认可保育所的招收定员数除以0～5岁儿童人口数。

③ 卡德等人的论文曾遭到加州大学欧文分校的大卫·纽马克（David Neumark）等人的反驳。此外，目前仍有很多研究在进行探讨，尚未得出最终结论。尤其此处介绍的论文

结论是否适合日本国情还有待进一步探讨。日本的相关实证研究可参考大竹文雄、川口大司、鹤光太郎合编的书以及鹤光太郎的相关文章（详见文末参考文献）。

第5章　看电视会导致孩子学习能力下降吗?利用第三变量的工具变量法

广告费打折，该怎样利用

本章将介绍“工具变量法”。

让我们再回到珠宝店的事例。你想评估广告对销售额有没有因果效应。安排下属调查后，遗憾地发现投放广告的店铺和不投放广告的店铺很难被视为可比较的组。因为投放广告与否很大程度上是由各店铺的店长决定的，投放广告的店铺与不投放广告的店铺相比，店长的干劲可谓天差地别。这种情况下，广告与销售额之间可能实际上不存在因果关系，但“店长的干劲”这个混杂因素使二者间显现出“伪相关”。

就在对着销售额数据一筹莫展时，你注意到了一件事。几年前，某地方报社一时起意，搞了一次广告费打折活动。只要广告费用下降，该地区的店铺投放广告的可能性就会大幅提升。于是你想到了一个主意：能否利用这家报社的广告费打折活动，弄清楚广告对销售额的因果效应呢?

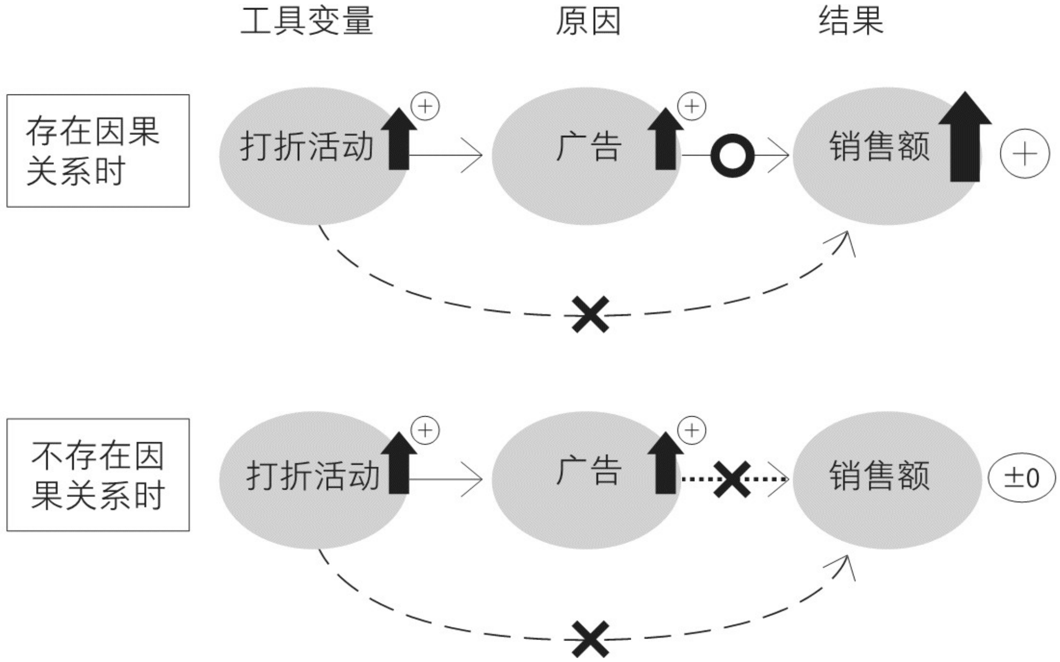
这就是工具变量法的基本思路。所谓工具变量，是指“对结果没有直接影响，但会通过影响原因间接影响结果”的第三变量。拿广告费打折的例子来说，这个变量虽然不会直接影响销售额，但会影响店铺是否投放广告的决策，从而间接对销售额产生影响。报社搞不搞广告费打折

活动与店长的干劲无关，也不会直接影响珠宝的销售额，但它会影响店铺是否投放广告的决策，从而间接影响到销售额。

那么为什么工具变量能帮助我们明确广告和销售额之间的因果关系呢?

先来看图表5-1。报社广告费打折活动开展后，投放广告的店铺自然会增多。不过，这个活动并不会对珠宝的销售额产生直接影响。

图表5-1　利用“工具变量”测量因果效应

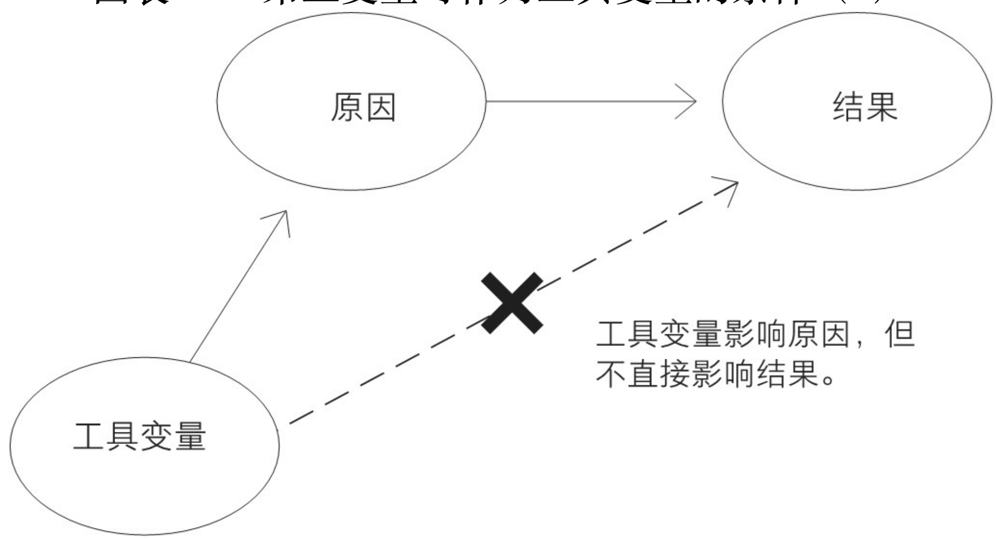


因此，如果广告与销售额之间存在因果关系，那么活动促使投放广告的店铺增多，就会导致销售额有所增长。相反，如果不存在因果关系，那么即使打折活动促使投放广告的店铺增多，销售额也不一定会上涨。

工具变量法成立的两个前提条件

工具变量需要满足两个条件。第一，工具变量要能够影响原因，但不直接影响结果。来看图表5-2。图中箭头表示原因与结果之间的因果关系，如“原因→结果”。箭头表示两者之间存在因果关系，其起点是原因，终点是结果。

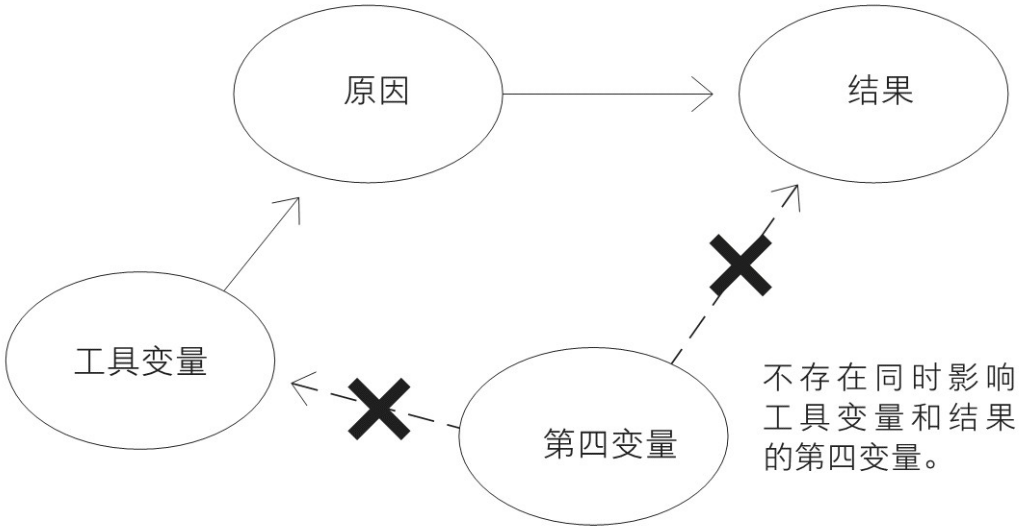
图表5-2　第三变量可作为工具变量的条件（1）



此时，要使工具变量成立，必须满足的条件是：“工具变量→原因”成立，且“工具变量→结果”不成立。

第二个条件如图表5-3所示，必须确保不存在同时影响工具变量和结果的“第四变量”。

图表5-3　第三变量可作为工具变量的条件（2）



“看电视”与“学习能力”是因果关系吗?

下面就来介绍一项应用了工具变量的研究。

不少家长担心孩子看电视时间过长。日本厚生劳动省的统计显示，小学六年级学生周一到周五每天约有2.2小时、周末每天约有2.4小时在电视机前度过。花这么多时间看电视，父母担心也在所难免。

很多人坚信电视会对孩子的成长、健康以及学习能力带来不良影响。美国俗语甚至把电视称作“傻瓜盒子”（idiot box）。

不过，看电视真的会降低孩子的学习能力吗？对这个问题，我们也需要慎重思考看电视与孩子学习能力的关系是因果关系还是相关关系。究竟是“因为看了电视，所以学习能力下降”（因果关系），还是“学习能力低下的孩子看电视时间更长”（相关关系）？

斯坦福大学的马修·根茨科（Matthew Gentzkow）等人研究了这项课题。他们关注了在1948年至1952年期间，美国以解决信号故障为由，对

新电视台的执照许可实行了长达4年的冻结令这一历史事件。

20世纪40年代到50年代中期，电视在美国得到普及，越来越多的家庭能够收看电视。然而，随着1948年新电视台执照许可叫停，生活在没有电视台地区的人们直到1952年解除冻结令之前，一直没有电视可看。也就是说，电视台执照许可冻结令分离出了1948年前就能收看电视的家庭（干预组）和在1952年前没能看上电视的家庭（对照组）。根茨科等人的研究发现，在电视普及初期的这段时间，家里有电视的孩子沉迷于电视，平均每天待在电视机前的时间长达3.5小时。

根茨科将“1948年到1952年间是否拥有电视”作为收看电视的工具变量。这一时期是否拥有电视很大程度上取决于调查对象是否居住在1948年前就已经取得执照的电视台的覆盖范围内。换句话说，电视台执照的相关规定明显会对看电视时间产生影响，但又不会直接影响孩子的学习能力，因此可以视作工具变量。

看电视可以提高偏差值

根茨科等人的分析结果出人意料：20世纪40年代至50年代前期，年幼时经常看电视的孩子在升入小学后的学力测试中，偏差值高出0.02。此外，也没有发现看电视对做作业的时间、升学志愿等方面产生不良影响。分析还指出，在英语不是母语，或者母亲学历较低，或者非白人人种的孩子当中，看电视还对成绩提高产生了显著效果。

根茨科等人认为，很多家长之所以不愿意让孩子看电视，或许是考虑到看电视会让孩子处于被动状态，夺走他们绘画或者运动的机会。确实，在拥有很多类似活动选项的富裕家庭里，电视的正面因果效应会很小，有时甚至会转化为负面效应。不过虽说如此，鉴于也有研究强调电视对贫困家庭孩子的积极影响，根茨科等人呼吁政府和教育人士不要毫无根据地大肆宣传电视的负面影响。

“母亲的学历”与“孩子的健康”是因果关系吗?

哥伦比亚大学的珍妮特·克里（Janet Currie）等人的研究也应用了工具变量法。经济学家克里曾通过多项研究证实幼年时孩子的健康状况会对长大后的学历及收入产生显著影响。而她接下来试图论证的是“幼年时孩子的健康状况取决于什么”。

克里认为母亲学历高或许会对孩子的健康状况产生积极影响，于是采用美国两百多万名母亲的数据进行了研究。

母亲学历高真的对孩子健康状况有好处吗？回答这个问题同样需要慎重思考母亲的学历和孩子的健康之间是因果关系还是相关关系。究竟是“因为母亲学历高，所以孩子很健康”（因果关系），还是“能获得高学历的母亲的孩子很健康”（相关关系）呢?

克里等人采用“17岁时的家庭居住地到大学的距离”作为工具变量。17岁时的家庭居住地附近有没有大学会影响上学或寄宿的费用，进而影响到是否要考大学的决策。克里等人的数据显示，17岁时的家庭居住地附近每多一所大学，该地区女性的大学升学率就会增长19%。此外，17岁时的家庭居住地到大学的距离不会直接影响多年后出生的孩子的健康状况，因此可以视作工具变量。

母亲上过大学，孩子更健康

克里等人的研究显示，本科以上高学历的母亲生育早产儿或低出生体重儿的概率更低，婴儿出生时的健康状况更好。究其原因，本科以上高学历母亲在怀孕期间吸烟的概率较低，接受产检的概率也更高。也就是说，上大学使她们养成了利于孩子健康的好习惯。

克里等人的研究让人们关注到教育带来的好处。芝加哥大学的迈克尔·格林斯通（Michael Greenstone）等人通过研究发现，股票、债券等金融投资所能获得的平均收益远不及上大学所能获得的收益，人们很难找到比让自己接受高等教育更有利的投资项目。此外，根据克里等人的研究，教育不仅对受教育者本人有益，下一代也会从中受益。可以说，对社会来说，教育是回报最高的投资之一。

第5章的关键词

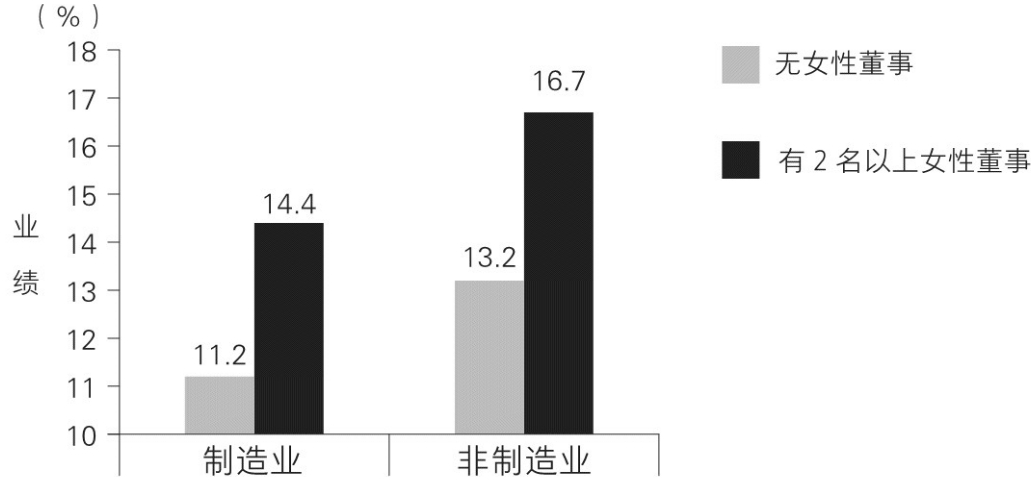
工具变量法

该方法利用“只能通过影响原因而间接影响结果”的工具变量，使受到干预的组（干预组）和不受干预的组（对照组）形成可比较的状态。不过，这种方法只有在满足两个前提条件时才能发挥作用。第一个条件是，工具变量会影响原因，但不会直接影响结果。第二个条件是，不存在同时影响工具变量和结果的第四变量。

COLUMN 5
增加女性管理层成员能促进企业成长吗?

2016年，安倍内阁出台“女性活跃推进法”。这项法案作为成长战略的一环，为女性管理者的比例设置了数据目标。按照这个规定，企业和地方政府必须制订并公开实现女性管理者比例目标的行动计划。某项调查结果显示，女性董事人数较多的企业业绩更好（图表5-4），至于二者之间是因果关系还是相关关系，还有待更严谨的论证。

图表5-4　女性董事人数多的企业业绩更好吗?

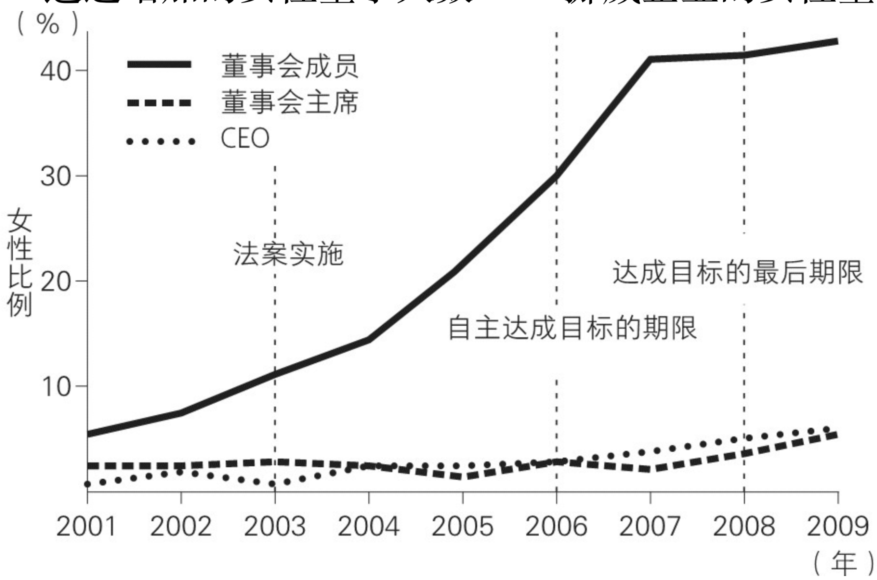


（注）企业业绩用经常性ROE（净资产回报率，Return On Equity）衡量。该指标体现企业的盈利能力，由经常利润除以资产净值得出。

（数据来源）日兴财务智囊机构《2014 年度产业经济研究外包项目（关于企业女性活跃及发展促进情况的调查）报告》。

挪威曾出台过一项很有冲击力的法案：企业必须在2008年前将女性董事的比例提高至40%，否则将面临解散。南加州大学的肯尼斯·R.埃亨（Kenneth R. Ahern）等人利用这一事件分析了女性董事比例与企业价值之间是否存在因果关系。①结果如图表5-5所示，在这项法案开始实施的2003年，挪威上市公司的女性董事比例不足10%，而且各企业之间差距也很大。

图表5-5　迅速增加的女性董事人数 ——挪威企业的女性董事比例



（数据来源）本图表由笔者根据埃亨和迪特马尔（Ahern and Dittmar）的研究（2012）制成。

得知这一情况后，埃亨等人决定将“法案实施前各企业的女性董事比例”作为工具变量。简单来说，法案实施前女性董事比例已经很高的企业在法案实施后，不难完成将比例提高至40%的目标，因此在2003年到2008年间，女性董事人数应该不会显著增加。相比之下，法案实施前女性董事比例较低的企业则会在同一时期迅速增加女性董事的人数。这样一来，“法案实施前的女性董事比例”会对2003年到2008年间各企业的女性董事人数增长率产生影响，而且，该变量应该不会对目前的企业价值产生直接影响，因此可以视作工具变量。

埃亨等人得出的结果出乎意料：女性董事比例的提高会降低企业价值。具体来说，女性董事比例每增加10%，企业价值会减少12.4%。强制设定女性董事比例的数值目标，会大幅降低企业价值，使股东担负相应的成本。

为什么会这样呢？深入调查发现，这一期间就任董事的女性大多比原有董事年轻，没有董事经验，或者多来自其他行业。不仅如

此，其中不少人的姓氏和原有董事相同。也就是说，她们很可能是原有董事的妻子或女儿。

为了完成政府制定的女性董事比例目标，很多挪威企业不得不把一些经验尚浅、缺乏管理者素质的女性推上董事职位，以解燃眉之急。这种做法导致了企业价值下降。

为避免误会，在此需要强调，我们介绍这项研究的目的并非反对提高企业中管理层女性成员的比例。公平且多样化的社会理应受到支持。挪威政府制定女性董事比例目标的初衷也是尊重男女就业机会均等的精神，建设更加公平的社会。

但是，仅通过制定硬性指标来增加女性管理层成员，以达到鼓励女性参与管理、提高企业价值的目的，有时反而会适得其反。挪威的事例已成为他国可资借鉴的教训。②

日本正在推行的女性活跃推进政策也应避免单纯制定数字目标的做法，而应通过提高工作形式的灵活性、构建没有性别歧视的公平的考核及待遇制度等措施，打造女性管理者自然增加的工作环境。

① 埃亨等人用“托宾Q值”作为代表企业价值的变量，通过负债和股价的市值之和除以资产市值求得。

② 在管理学领域，人才多元化分为女性、外国人等可以根据外表判断的“特征型”人才多元化，和注重实际工作能力和经验的“工作型”人才多元化。一项综合了以往研究的元分析显示，对提高企业价值具有重要意义的是后者，即工作型人才多元化。

第6章　和学霸做朋友，学习能力会提高吗?关注跳跃的断点回归设计

利用“49人店铺”与“50人店铺”的差异

本章将介绍“断点回归设计”。让我们继续以珠宝店为例。就职于珠宝店的你还在为如何论证广告的效果而一筹莫展。这时，下属向你提出了一个方案：只有员工人数50人以上（含50人）的大型店铺才在年底圣诞季商战期间投放广告。

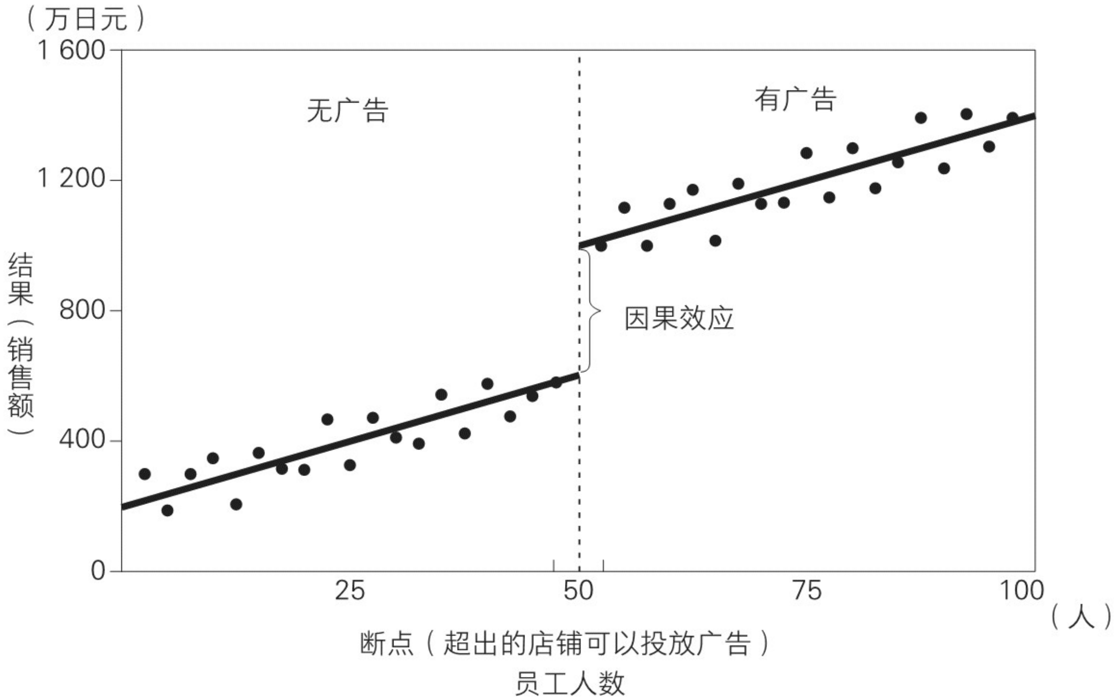
对此，你产生了一个想法：按照这个方案，员工人数49人的店铺不投放广告，而50人的店铺投放广告。不过，49人的店铺和50人的店铺除了是否投放广告之外，应该没有其他特别显著的差异。这样的话，只要比较49人店铺和50人店铺的销售额，不就可以明确广告对销售额的因果效应吗？其实，采用“断点回归设计”就能实现这个想法。

图表6-1体现了员工人数和各店铺销售额的关系。当然，销售额和员工人数呈正比趋势。按照下属的方案，只有员工人数50人以上（含50人）的店铺在圣诞季投放广告。这种情况下，“员工人数50人以上（含50人）”这个条件是一个没有什么特殊理由、任意设置的数值（我们称其为“断点”）。和你想象的一样，断点（员工人数50人）上下的店铺销售额确实没有显著差异。即员工人数刚满50人并因此能够投放广告的店

铺（干预组）和员工人数差那么一点不到50人并因此未能投放广告的店铺（对照组）可以视作可比较的组。

从图表6-1可以发现，销售额在断点处出现了大幅“跳跃”。断点出现的“跳跃”是投放广告导致的，因此，“跳跃”的幅度即可视为投放广告对销售额带来的因果效应。

图表6-1　断点前后出现“跳跃”——断点回归设计的事例



断点回归设计成立的前提条件

用断点回归设计推算因果效应，必须满足一个重要的前提条件：在连续变量的断点附近，没有发生过影响结果的其他事件。比如，假设员工人数50人以上（含50人）的店铺不仅可以投放广告，还会根据销售额

发放奖金。在这种情况下，就无法判断断点的“跳跃”是广告的效果还是奖金的效果了。

“朋友的学力”与“自己的学力”是因果关系吗?

下面介绍几项应用断点回归设计的研究成果。每当备考季节临近，总有不少家长祈祷自家孩子能够超常发挥，考入偏差值高的学校。他们可能认为在偏差值高的学校就读的肯定都是学习能力强的学生，即使自家孩子的实力可能稍有不足，但能和学霸朋友们一起学习，孩子的学习能力自然也会随之提高。可能正因为他们的这种想法，“学生学习能力高”的学校附近房价和地价才会连年攀升吧。

经济学领域将朋友给予的影响称为“同群效应”。家长们的想法相当于认为同群效应会给自家孩子的学习能力带来积极影响。然而，这一观点有待严谨的论证。到底是“因为和学霸做了朋友，所以自己的学习能力提高了”（因果关系），还是“学习能力强的孩子更愿意与学霸做朋友”（相关关系）呢?

麻省理工学院的安格瑞斯特（Joshua Angrist）等人研究了这一课题。波士顿和纽约各有三所特殊的公立高中，专门面向备考大学的学生。不过，这些学校和日本一样，也有入学考试，只有合格者才能入学，即所谓的“精英高中”。

在精英高中的入学考试中落榜的学生只能去上其他公立高中。和通过入学考试选拔优秀生源的精英高中相比，其他公立学校的学生平均学习能力自然要低得多。

安格瑞斯特等人认为，勉强通过合格线考入精英高中的学生（干预组）和差那么一点没到合格线只得去上其他高中的学生（对照组）属于可比较组。他们试图利用这种情况，将合格分数作为断点值，运用断点

回归设计论证和学习能力强的朋友一起度过高中生活能否对学生的学习能力产生因果效应。

和学霸在一起，也无法提高自己的学习能力

安格瑞斯特等人的研究结果显示，在波士顿和纽约的所有学校，断点前后并未观测到学习能力的“跳跃”现象。

关于“同群效应”是否存在尚无定论，但也有不少探讨和学习能力强的人做朋友对学习能力的因果效应的研究得出了和安格瑞斯特等人相同的结论。

例如，美国全国经济研究局（NBER）的杰弗里·克林（Jeffrey R. Kling）等人对美国政府实施的一项大型随机对照试验“希望迁移计划（Moving to Opportunity）”进行了研究。这项计划从贫困家庭中随机抽选一些家庭，向入选家庭发放能够搬迁到贫困率较低区域的优惠券。

入选家庭的孩子们在搬迁到新家后，会和比自己学习能力强的朋友们一起度过校园生活。但是，他们和继续生活在原来地区的落选家庭的孩子们的学习能力差异在统计学上并不具有显著性。

很遗憾，现实辜负了众多家长的期望，和学霸一起度过高中生活对自家孩子的学习能力几乎不会产生任何影响。这个结果或许说明，不应该过分期待周围人的影响而忽略了提高自己的实力。

“医疗费用自付比例”与“死亡率”是因果关系吗?

再来介绍一项应用了断点回归设计的研究。在第2章中，我们探讨了医疗费用自付比例的问题。一些观点建议通过提高自付比例来控制医疗费用的增加，也有人担忧这样做会迫使老年人减少就医次数，对其健康状况造成不利影响。针对这个课题，加拿大西蒙弗雷泽大学的重冈仁利用日本的数据进行了研究。

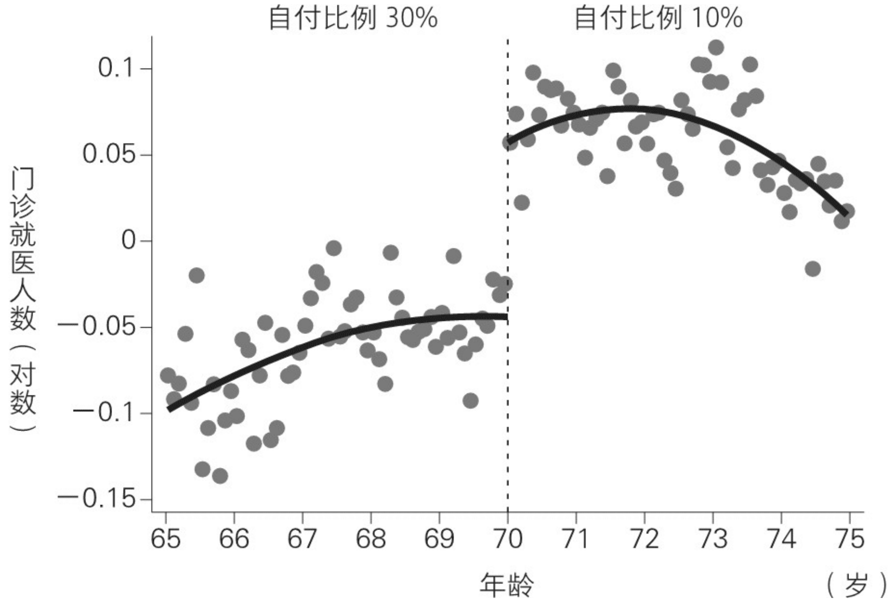
重冈着眼于日本老年人在年满70岁以后医疗费用自付比例从30%降至10%的制度，将70岁作为断点值，试图研究断点前后老年人在健康状况及医疗服务使用频率上的变化。①也就是说，他认为70岁0个月的人（干预组）和69岁11个月的人（对照组）属于可比较组。

第2章介绍的兰德健康保险实验（参考正文第36页）得出的结论是，提高自付比例对健康状况没有影响。然而在美国，由于人们年满65岁就会自动加入国家公共医疗保险（联邦医疗保险，Medicare），所以兰德健康保险实验中未包括65岁以上的老年人。重冈这项研究的卓越之处在于，对兰德健康保险实验未能覆盖到的老年人群体自付比例与健康和死亡率之间的因果关系做了论证。

提高老年人医疗费用的自付比例对死亡率没有影响

重冈的研究结果显示，老年人的自付比例下降后，门诊就医患者数量增加了10.3%。（见图表6-2）也就是说，在断点值70岁前后观测到了门诊就医患者数量和住院患者数量的“跳跃”。②特别是膝盖疼痛等关节痛患者的数量增长十分明显。

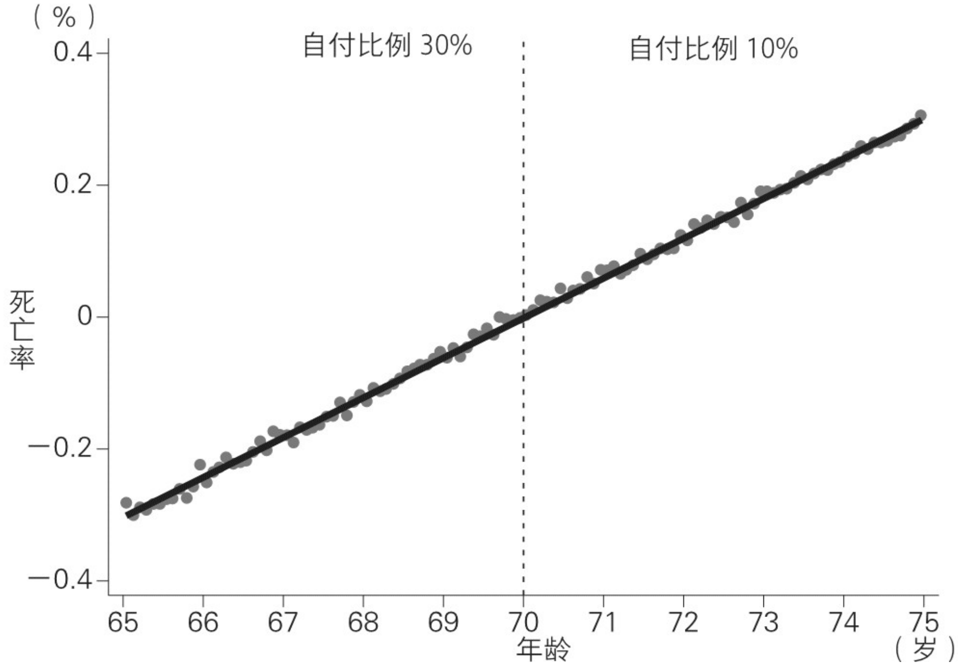
图表6-2　门诊就医患者人数随自付比例的降低而增加



（数据来源）重冈（2014）

那么，自付比例的降低对健康和死亡率造成了怎样的影响呢？重冈得出的结论出人意料。如图表6-3所示，在断点值70岁前后，并未观测到死亡率的“跳跃”。也就是说，即使自付比例降低，就医或住院频率增高，死亡率也没有发生变化。不仅如此，在调查中回答身心健康以及自我健康状况良好的受访者的比例也未出现“跳跃”。

图表6-3　改变自付比例，死亡率不变



（数据来源）重冈（2014）

这说明，医疗费用自付比例降低后，虽然老年人去医院的次数变多了，但这对他们的死亡率和健康状况并不会产生影响。这项研究有望为医疗制度改革铺路，在不损害老年人健康的前提下削减医疗支出。③

第6章的关键词

断点回归设计

断点回归设计是利用任意决定的断点值两侧自然成为干预组与对照组的情况，推算因果效应的方法。断点回归设计成立的前提条件是断点值周围没有发生影响结果的其他事件。

COLUMN 6
“激素替代疗法”的陷阱

前面曾提到随机对照试验的证据等级比较高。下面就来介绍一项研究，它用随机对照试验完全逆转了用观测数据得出的结论，也因此而让“激素替代疗法”受到了广泛关注。

众所周知，女性患心肌梗死等动脉硬化导致的疾病的风险要低于男性。但女性绝经后，患心肌梗死的概率会增至和男性同等的水平。因此有假说推测是雌激素降低了动脉硬化的风险。

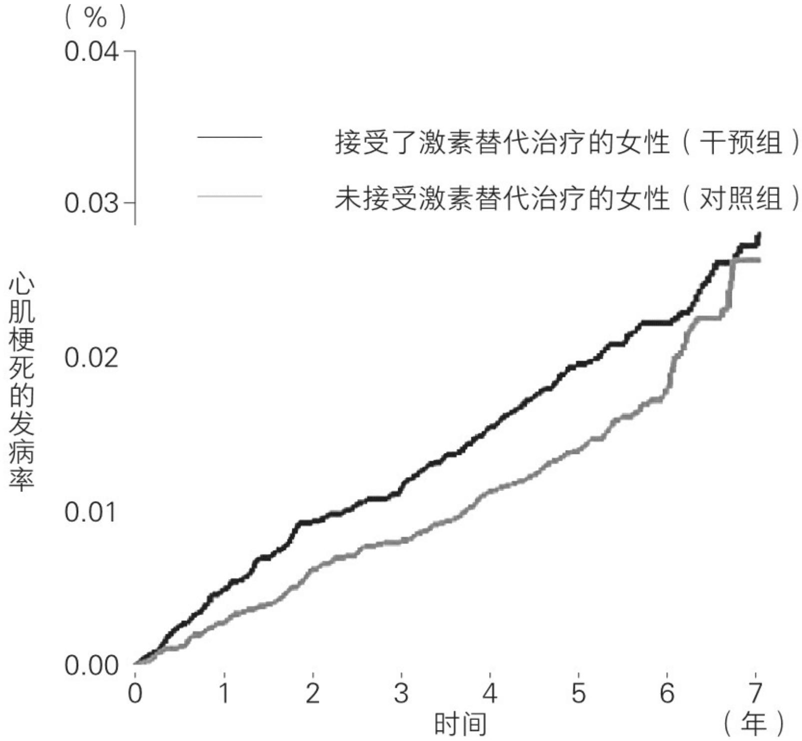
于是人们开始通过“激素替代疗法”，为绝经后的女性补充雌激素，试图由此降低心肌梗死的患病风险。针对激素替代疗法的初期研究（使用观测数据的研究）均得出了相同结论：持续接受激素替代疗法的患者患心肌梗死的风险更低。激素替代疗法不仅能够预防心肌梗死，还具有缓解更年期症状的功效，因此受到重视健康的女性的广泛支持。

20世纪90年代后期，为了重新确认激素替代疗法的效果，人们进行了一次随机对照试验，将已绝经的女性随机分成接受激素替代治疗组（干预组）和不接受治疗组（对照组），进行了比较。

该随机对照试验得出了令人意想不到的结果。在实验开始了5年之后的2002年，人们发现接受了激素替代治疗的女性比未接受该治疗的女性乳腺癌发病率更高，且这一差异在统计学上具有显著性。该项使用激素替代疗法的研究由于会对受试者的健康带来危害，所以很快就被叫停。

即便如此，仍然有很多人期待该疗法能够降低心肌梗死的患病风险。然而，之后的分析发现，接受了激素替代治疗的女性患心肌梗死的风险反而更高。（见图表6-4）

图表6-4　随机对照试验和观测研究的结果完全相悖——激素疗法与心肌梗死的关系



（注）心肌梗死的发病率用累计风险率体现。

（数据来源）J. E. 曼森（J. E. Manson）等（2003）。

那么为什么会出现这种情况呢？因为当年接受激素替代治疗的主要是教育或收入水平较高的女性，她们原本就很注重健康、饮食、运动等日常生活习惯，因此她们患心肌梗死的风险本来就比较低。随机对照试验之前的分析可能没有讨论诸如“对健康的关注程度”等混杂因素，从而误把伪相关当作了因果关系。这个事例使人们再次意识到随机对照试验在因果推理中的重要性。

① 2017年2月的政策为70～74岁老年人的医疗费用自付比例为20%，75岁以上为自付10%。该研究所使用的数据在统计时的政策为70岁以上自付10%。

② 该研究发现，医疗费用自付比例增加10%，患者对医疗服务的需求就会下降约2%。这一需求变化与前文中兰德健康保险实验得出的结果几乎相同。

③ 需要注意的一点是，由于重冈的研究只讨论了短期的死亡率，所以尚不清楚改变医疗费用自付比例是否会对老年人的长期健康状况产生影响。尤其是正如兰德健康保险实验显示的，对贫困人群来说，自付比例的提高可能会对健康状况产生不利影响。

第7章　上好大学收入就会更高吗？组合相似个体的匹配法

找出相似的店铺

本章介绍最后一种准实验方法——“匹配法”。我们依旧用珠宝店的事例来思考。假设你还在为投放广告的店铺和不投放广告的店铺不可比较一事发愁。

这时，你突然产生了一个想法。你所在的珠宝店目前共有100家店铺。其中30家投放了广告，70家未投放广告。从未投放广告的70家店铺中，选出和投放广告的30家店铺情况相似的30家店铺，将二者进行比较不就好了吗？这便是匹配法的思路。

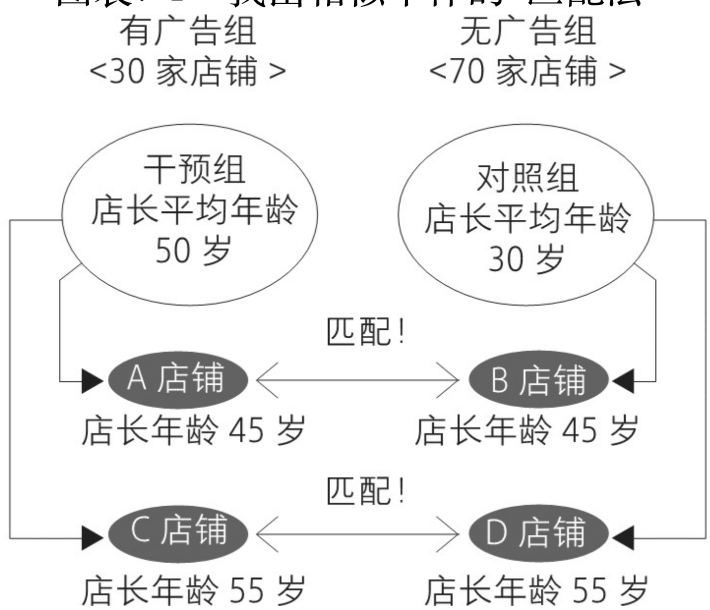
所谓匹配法，即从对照组中选出和干预组非常接近的配对，使两组成为可比较组的方法。

如第5章所述，投放广告与否很大程度上取决于各店铺店长的判断。投放广告的店铺，店长大多年纪比较大，从业时间更长。假设他们在其他方面没有明显差异。因此，投放广告的30家店铺的店长平均年龄是50岁，而不投放广告的店铺，店长平均年龄是30岁。

假设某家投放广告店铺的店长现年45岁。按照匹配法的思路，先从未投放广告的店铺中选出店长同为45岁的店铺进行匹配。假设另一

家投放广告店铺的店长现年55岁，那么就用相同的方法选出店长同为55岁的无广告投放店铺进行匹配。这样一来，最终便可以形成店长平均年龄均为50岁的干预组和对照组，这两个组为可比较组。（见图表7-1）这种情况下，相当于店长“年龄”的变量称为“协变量”①。

图表7-1　找出相似个体的“匹配法”



整合多个协变量的“倾向得分匹配法”

协变量未必只有一个。如果只考虑店长年龄相同的店铺，那么从对照组中找到符合这个条件的店铺进行匹配并不难。然而，如果存在多个协变量，恐怕就不一定能从对照组中找到所有条件都完全一样的店铺了。

这种情况下，我们可以使用匹配法中的“倾向得分匹配法”（Propensity Score Matching）。“倾向得分匹配法”将多个协变量整

合成一个得分，用这个得分进行匹配。倾向得分指 “分入干预组的概率”。例如，考虑多个协变量后，如果某店铺投放广告的概率为50%，则该店铺的倾向得分就是0.5；如果某店铺投放广告的概率为30%，则该店铺的倾向得分为0.3。

如果投放广告的店铺（干预组）中有某店铺倾向得分为0.5，那么就从未投放广告的店铺（对照组）中选出倾向得分同为0.5的店铺进行匹配；如果某店铺倾向得分为0.3，那么就从对照组选出得分同为0.3的店铺进行匹配。对采用这种方法最终形成的两个组进行比较，可以发现用于计算倾向得分的所有协变量（平均之后）在两个组具有相同的值。

为了便于读者理解倾向得分匹配法的整体概念，下面就用笔者自制的数据来推算广告对销售额的因果效应。假设影响销售额的因素包括店长年龄、店铺所在地区的人口以及店铺所在地区的人均收入（“店长年龄”“地区人口”和“地区人均收入”等3个变量为协变量）。

来看图表7-2。如果只看这一张表，似乎不投放广告的店铺的销售额反而更高。不过事实是否真是这样呢？你所在的珠宝店不应该投放广告吗?

图表7-2　无法对“协变量”各不相同的两个组进行比较

		有广告 （30 家店铺 / 干预组）	无广告 （70 家店铺 / 对照组）
协变量	店长年龄	50 岁	30 岁
	地区人口	30 万人	100 万人
	地区人均收入	450 万日元	650 万日元
结果	销售额	1 000 万日元	1 400 万日元

显然，投放广告的干预组和未投放广告的对照组不是可比较组。投放广告的店铺店长年龄比较大，也许是公司特意把经验丰富的店长

派到了销售额较低的地区。如果单纯从有无投放广告的角度对销售额进行比较，很可能得出广告对销售额具有负面效果的错误结论。

这种情况就正好可以用到倾向得分匹配法。用店长年龄、地区人口、地区人均收入这三个协变量计算出倾向得分。然后，从干预组和对照组中依次找到倾向得分相近的店铺进行匹配。这样便可以得到干预组的30家店铺和对照组的30家店铺。

来看图表7-3。这是进行倾向得分匹配后的干预组和对照组的对比。匹配之后，只要协变量的分布一致（经济学家们将这种现象称为“协变量达到平衡”），我们就可以说这两个组是可比较的。

图表7-3　利用倾向得分匹配法使“协变量”达到平衡

		有广告 （30 家店铺 / 干预组）	无广告 （30家店铺/对照组）
协变量	店长年龄	48 岁	47 岁
	地区人口	40 万人	41 万人
	地区人均收入	460 万日元	450 万日元
结果	销售额	1 200 万日元	1 000 万日元

200 万日元的因果效应

达到了这个状态，我们才可以说构造出了两个可比较组：考虑多种属性，两家店铺投放广告的可能性相同，只是碰巧其中一家投放了广告，而另一家没投放广告。也就是说，这种情况下，对照组才能代表“如果干预组没有投放广告”的反事实，我们也才能明确干预组和对照组的销售额差异200万日元（1 200万日元-1 000万日元）是广告带来的因果效应。

倾向得分匹配法成立的前提条件

倾向得分匹配法成立必须满足两个前提条件。第一个前提条件是，所有会影响结果的协变量均为可转化成数值的数据。有些情况下也会存在无法转化成数值的协变量[2]。在珠宝店的例子中，“附近居民对冠婚葬祭仪式的重视程度”即属于这种协变量。如果人们对冠婚葬祭仪式的重视程度高，他们在珠宝店的消费一般也会更多，不过这种变量通常很难转化为数值。

第二个前提条件是所有协变量都必须用来计算倾向得分。

“大学偏差值”与“收入”是因果关系吗?

下面介绍普林斯顿大学的克鲁格等人撰写的两篇论文，是运用匹配法进行的出色研究。

日本被称为学历社会，但即使都是大学毕业生，之后的人生也并不会相同。一桥大学的神林龙等人的研究显示，20世纪90年代大学毕业生和高中毕业生的工资差距并未拉大，但大学毕业生之间的工资差距却有所增加。

那么同为大学毕业生，上了偏差值高的大学，将来的收入就会更高吗? [3]对此也需要慎重思考。是“因为上了偏差值高的大学，所以收入更高”（因果关系），抑或只是“潜在能力强、以后会进入高收入职业的人选择了偏差值高的大学”（相关关系）呢?

克鲁格等人尝试用匹配法解答这个问题。首先需要介绍的是，美国的大学入学考试和日本略有不同。日本的大学入学考试以笔试为

主，而美国除了笔试成绩，还会综合考察高中成绩、教师推荐信、报考理由等因素进行选拔。

克鲁格等人采用每名考生“考上的大学和未考上的大学”进行了匹配。例如假设有两名考生考上了A大学和B大学，但没有考上C大学。他们考上和没考上的大学均相同，因此至少从用于判定大学入学考试合格与否的信息（高中成绩、教师推荐信、报考理由等）来看，二人是可比较的。④

然后比较二人考上的两所大学，大学A的偏差值要比大学B更高。假设其中一人选择就读大学A，而另一人选择了专业对自己更有吸引力的位于当地的大学B。那么只要对这两个人进行比较，就能推算出就读偏差值高的大学对未来收入的因果效应。

上偏差值高的大学，收入也不会增长

匹配法的分析结果出乎意料。在考上某大学且就读于该大学的学生组（干预组）和同样考上该大学但最终选择了偏差值较低的其他大学的学生组（对照组）之间，毕业后的工资差异在统计学上不具有显著性。不少人相信就读偏差值高的大学便能获得更好的收入，而就克鲁格等人的研究结果来看，这种因果关系实际上并不存在。不过，该结论不适用于非裔美国人等少数群体以及父母受教育程度均不到大学水平的贫困家庭。对此，克鲁格等人给出的解释是：在偏差值高的大学建立的人际社交网络或许会有利于少数人种群体或贫困阶层的发展。

克鲁格等人的研究运用了匹配法，此外也有运用倾向得分匹配法进行的研究。芝加哥大学的丹·布莱克（Dan Black）等人采用美国的数据推算了大学偏差值对未来收入的因果效应。

布莱克等人计算倾向得分时列出的协变量包括年龄、人种、出生地、学力测试的成绩、毕业高中的规模、毕业高中的教师素质、父母的学历和职业、童年有无订阅报纸或在图书馆借阅的经历等。用倾向得分匹配法进行匹配之后，他们得出了和克鲁格等人相同的结论：就读于高偏差值大学的组（干预组）与就读于低偏差值大学的组（对照组）毕业后的收入差异在统计学上不具有显著性。

正如克鲁格等人所说，“并非对所有学生来说，就读于偏差值高的大学都意味着未来收入最大化”，“世界上没有任何一所大学能够保证毕业生将来会获得更高的收入”。

不仅如此，克鲁格等人在论文中还援引了明尼苏达州诺斯菲尔德的名门私立大学嘉尔顿大学校长斯蒂芬·R.刘易斯（Stephen R. Lewis）的话。被问及如何看待大学排名时，刘易斯答道：“问题不在于哪所大学是最棒的，而在于对谁来说是最棒的。”这句话可谓概括了这一系列研究的结论。

第7章的关键词

匹配法

该方法运用能够影响结果的协变量，从对照组中选出和干预组相似的样本进行匹配，并对两个组进行比较。存在多个协变量时，也可以将它们整合成一个得分进行匹配（倾向得分匹配法）。匹配法成立的条件是影响结果的所有协变量均可观测。

COLUMN 7
商务版随机对照试验“A/B测试”

随机对照试验在商业领域也逐渐得到了广泛应用。“A/B测试”就属于这种情况。

让我们再回到珠宝店的事例。热销产品的首饰盒现有两款候选设计A和B，你在犹豫应该选用哪一款。一名销售员断定设计A更符合公司的风格，另一名销售员则说其他公司的畅销产品都与设计B很像。即使选择了设计A的首饰盒，可能你也会情不自禁地想象反事实的情况：如果用了设计B会怎样（反之亦然）。

在这种情况下可以采用“A/B测试”。将所有店铺随机分成两组，一半店铺使用首饰盒A，余下的店铺使用首饰盒B。一段时间之后，再把销量更好的店铺所用的首饰盒推广到所有店铺即可。

A/B测试最常见的应用领域是购物网站。或许有人曾经注意到，即使是同一商品，不同地区或不同用户看到的图片、宣传图、宣传语等可能不尽相同。这种情况正是企业在进行A/B测试。

南非一家名为Credit Indemnity的金融机构所做的A/B测试，结果十分耐人寻味。这家金融机构向大约5万名客户随机发放了直邮广告。这些直邮广告的内容有多种不同形式，有些贷款方案比较详尽，有些则比较简略；有些广告中附有照片，有些没有照片。其目的在于调查直邮广告的内容对贷款申请数量会产生多大程度的影响。

结果发现，附有性感女郎照片的直邮广告具有促进男性客户申请贷款的功效。此外，只印有1个贷款金额和还款方案的直邮广告比写有4个方案的广告收到的申请更多。也就是说，越简洁、信息量越少的直邮广告，增加客户的效果反而越显著。这或许说明直观、简明易懂、让人看到就能马上做出答复的广告，效果会更好。

①也许有些读者想不通协变量和混杂因素有何区别。协变量指原因与结果以外的所有其他变量。比如在现有数据中，除了原因和结果的变量，其他所有变量都是协变量。而混杂因素是这些协变量中“同时影响原因与结果的变量”。也就是说，协变量中包括混杂因素，也包括非混杂因素。

②“无法转化成数值的协变量”包括原本就未测量的变量和测量过但因某些原因未能掌握的变量等。也就是说，其中也包括本来可以转化成数值，但由于某些原因无法以数值化数据的形式运用于分析的协变量。

③美国没有“偏差值”的概念，人们普遍接受的是被称为“SAT”的学术水平测验考试。SAT相当于可以多次参加的中心考试（即日本的全国统一大学入学考试），录取难度越高的大学要求的SAT得分也越高。因此，这些研究使用的是美国版“偏差值”，即各所大学录取学生的SAT平均分。

④克鲁格等人的研究的卓越之处在于，采用大学录取情况进行匹配，对以往倾向得分匹配法无法充分处理的“无法转化成数值的协变量”也做了处理。大学在考虑是否录取一名考生时，很可能会根据教师推荐信和报考理由等来考察学生的学习动力或潜力等无法转化成数值的信息。而使用录取情况进行匹配，则可以将原本无法转化成数值的协变量一同纳入考察范围。

第8章　便于分析现有数据的回归分析

假如现有数据都不适合用来评估因果关系……

就职于珠宝店的你盯着面前的数据。公司在一本半年前创刊的杂志上投放广告，今天你收到了投放广告后的销售数据。领导让你根据这些数据分析下一季度是否应该继续在该杂志上投放广告。然而遗憾的是，该杂志创刊没多久，无法用断点回归法、双重差分法等准实验方法进行分析。

当然，运用随机对照试验或准实验进行分析，正确评估因果效应的可能性会更大。然而，在实际工作和生活中，绝大部分数据只是对经济活动结果的记录，并非以分析为目的而收集的（多数“大数据”都是这样）。遇到这种情况，随机对照试验或准实验大多没有用武之地。

要不知难而退，告诉领导“从这些数据中看不出广告的效果”？恐怕没有几个人敢这么做。有时候就算无法做出全面分析，也必须得出一些结论。

那么像这样，我们掌握一些数据，但又不适合用来评估因果关系时，应该怎么办呢？

这种时候就要用到“回归分析”。关于这种分析方法，市面上可以找到很多专业书籍，因此这里不再赘述，只简单地介绍一下其精髓部分。

回归分析分为两种：一元回归分析和多元回归分析。一元回归分析是评估两个变量之间的关系的方法。但如果存在混杂因素，这种方法无法排除混杂因素的影响。

相比之下，多元回归分析可以排除混杂因素的影响，并在此基础上评估原因与结果的关系。假如我们掌握所有混杂因素的数据（不过这种情况极为罕见），用多元回归分析也完全能够证明因果关系。

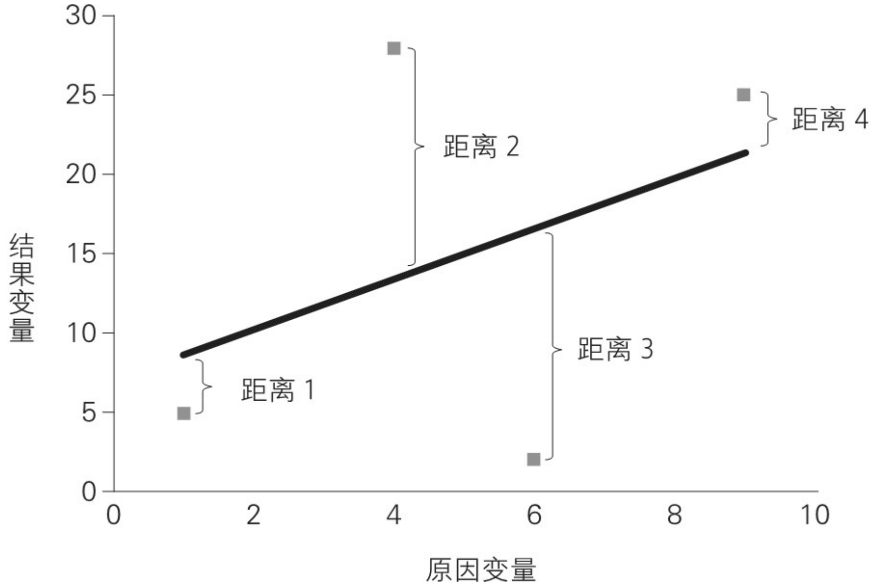
画一条表示数据的“回归线”

下面先用一元回归分析介绍回归分析的基本概念。假设有4个数据如图表8-1所示，每个数据都代表一种原因与结果的组合。你希望能确定这两个变量之间是否存在因果关系。

所谓回归分析，就是在这4个点之间画出一条“回归线”。如果能画出回归线，这条线的“倾斜程度”就代表了每增加一个单位的原因会给结果带来的变化。也就是说，回归线的倾斜程度便是原因对结果的“因果效应”。

下面讲解画回归线的具体方法。首先，在4个点之间大致画一条线。从这条线向各数据点（图表8-1中的方形）垂直画线，测量线到数据点的距离。我们用距离1表示第一个数据点和线之间的距离，用距离2表示第二个数据点和线之间的距离，那么可以得到从距离1到距离4这4个点到线的距离。使4个距离之和最小的线便是“回归线”。

图表8-1 “回归线”示意图



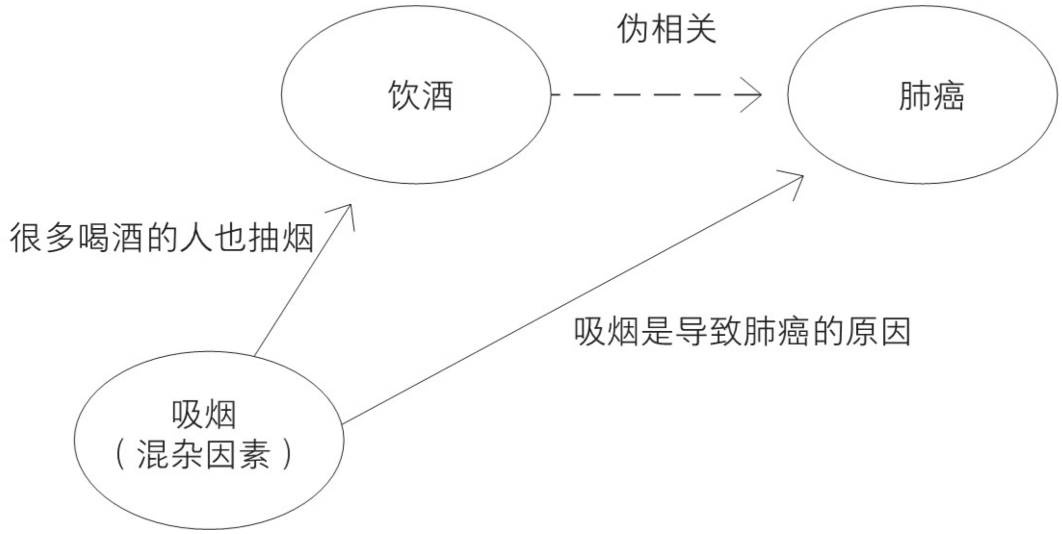
实际操作中无须手工画线，统计分析软件会帮助我们画出回归线。

用“多元回归分析”排除混杂因素的影响

如果想在一元回归分析中用回归线的倾斜程度代表因果效应，必须满足“不存在混杂因素”的前提条件。然而，现实中很少有情况符合这个条件。多元回归分析则可以排除混杂因素的影响。多元回归分析使混杂因素的值保持固定不变①，以便我们能够画出表示原因与结果的关系的“回归线”，并根据回归线的倾斜程度来评估因果效应。

下面我们以饮酒与肺癌的关系为例进行说明。假设我们现在怀疑饮酒与肺癌之间存在因果关系。因为众所周知，饮酒量越大，患肺癌的风险越高。这里我们需要注意一个问题：存在“吸烟”这个混杂因素。很多喝酒的人也抽烟，而吸烟也是导致肺癌的原因之一。（见图表8-2）

图表8-2　“饮酒”与“肺癌”之间存在混杂因素吗?



如果采用多元回归分析明确饮酒和肺癌之间是否存在因果关系，则需要在吸烟量相同的人之间比较饮酒量较多的人和饮酒量较少的人，看他们患肺癌的风险是否存在差异。这种做法便是“使吸烟量保持固定不变”，使用多元回归分析可以实现这一点。固定的吸烟量可以是1日0根（不吸烟的人），也可以是1日5根。这样就可以在排除了吸烟量这个混杂因素的影响的基础上评估饮酒与肺癌的关系了。②

多元回归分析的具体操作方法可以参阅专业书籍，使用一般的统计分析软件，任何人都能轻松进行多元回归分析。在会上被问及广告的效果时，如果只是单纯地比较目前和去年的数据，把广告与销售额之间的相关关系当作因果关系来看待的话，就有可能会被旁人指出混杂因素的存在，当众出洋相。

而如果能提前收集混杂因素的数据，在排除其影响的基础上再评估原因与结果的关系，无论是领导还是客户，都肯定会更加信任你给出的数字。

COLUMN 8 因果推理的发展史

经济学及统计学领域的因果推理

因果推理在经济学领域的历史其实并不长。20世纪40年代，挪威奥斯陆大学的经济学者特里夫·哈维默（Trygve Magnus Haavelmo）在论文中使用了与反事实概念相类似的表述，但这一表述从严格意义上来说还算不上对反事实的准确定义。

进入20世纪90年代后，出现了一些较大变化。当时哈佛大学经济学院的著名计量经济学家吉多·因本斯（Guido Imbens）、麻省理工学院的劳动经济学家安格瑞斯特开始和唐纳德·鲁宾（参考正文第14页）合作，将“鲁宾因果模型”引入经济学。③

哈佛大学和麻省理工学院都位于美国马萨诸塞州剑桥市，地缘上的接近或许也是促成他们合作的一个原因。鲁宾和因本斯合著的《统计学、社会科学及生物医学领域中的因果推理导论》（*Causal Inference for Statistics and Biomedical Sciences: An Introduction*）以他们多年来在哈佛大学经济学院的课堂讲义为基础撰写而成，2015年出版后旋即被誉为“因果推理”领域最经典的教科书。

因果推理在经济学领域的历史不长是有原因的。因为流行病学和生物统计学可以进行“临床试验”等实验，但是经济学等社会科学领域则很难组织实验。以人类为对象的实验除了资金和伦理方面的问题之外，大多还会伴随政治层面的难题。这些因素使得因果推理一直无法在经济学领域得到普及。

然而，进入21世纪以后，经济学领域出现了一些新动向。芝加哥大学的实验经济学家约翰·A.李斯特（John A. List）以及由发展经

济学家们组成的麻省理工学院扶贫实验室（J-PAL）的学者们克服重重障碍，开始进行大规模社会实验。

扶贫实验室堪称“实施随机对照试验”的专业组织，所有研究均采用随机对照试验。他们以“把易受政治风向左右的政策变为有理有据的政策”为目标，成功提高了随机对照试验的地位，使其成为“政策评估的理想方式”。

在经济学中，根据因果推理来评估政策效果的研究领域叫作“政策评估”，近年来该领域正在快速形成体系。

流行病学领域的因果推理

除经济学和统计学之外，因果推理的思维方式在其他学科中也得到了发展。其中之一便是流行病学。如果说“医学”是以个体为对象，研究其患病原因和治疗方法，那么“流行病学”就是要研究集体的患病原因和预防方法。医生、护士等拥有医学背景知识的读者或许对流行病学领域的因果推理更为熟悉。

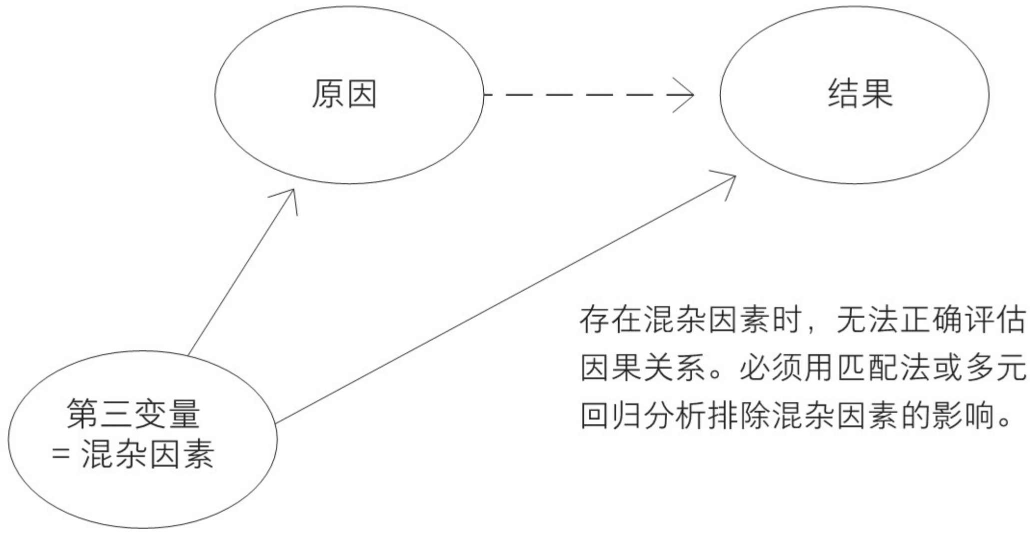
20世纪90年代中期，犹太裔的美国计算机科学家朱迪亚·珀尔（Judea Pearl）开发了运用有向无环图（DAG）来证明因果关系的方法。后来哈佛大学的詹姆斯·罗宾斯（James Robins）和米格尔·埃尔南（Miguel Hernán）、加州大学洛杉矶分校的桑德·格林兰（Sander Greenland）等人将有向无环图推广到医学及流行病学领域。最终，医学院和公共卫生研究院接纳了这种方法论并将其编入教材。

有向无环图的最大特点是用箭头示意图来体现因果关系。本书中多次出现的用箭头表达因果关系或相关关系的示意图也可算作有向无环图。

由于混杂因素（参考正文第7页）是“同时影响原因和结果的第三变量”，如果某一变量的箭头可以同时指向原因和结果，那么这

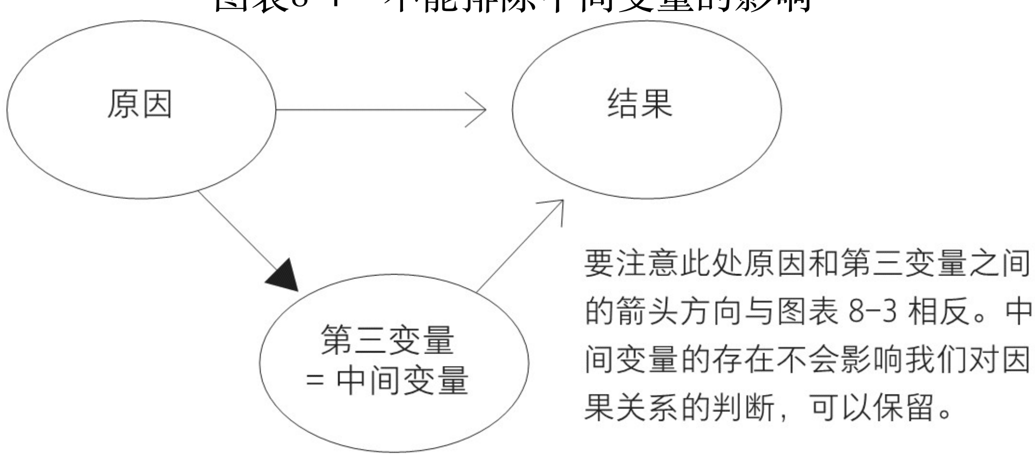
个第三变量便可以视为混杂因素。（见图表8-3）如果确定存在混杂因素，就必须用匹配法或多元回归分析排除其影响，否则无法正确评估因果关系。

图表8-3　必须排除混杂因素的影响



与此相对，如果箭头的方向相反，是从原因指向第三变量的话，则可以判定该变量不是混杂因素（该变量位于因果关系通道的中间，故名“中间变量”。见图表8-4）。如果用多元回归分析等方法处理中间变量，则会低估原因原本具有的影响。

图表8-4　不能排除中间变量的影响



① “保持固定不变”也可以叫作“校正”“控制”。

② 有研究显示，饮酒量处于中等程度及以下时，饮酒与肺癌之间没有关系，但就大量饮酒的人来说，饮酒与肺癌之间可能存在因果关系。也有研究使用日本人的数据，在论文中指出在吸烟者中发现饮酒与肺癌之间存在因果关系，但从整体上来看，这一问题尚未得出结论。

③ 鲁宾因果模型认为，所有人都有两种“潜在结果”，一种是受到干预后的潜在结果，另一种是未受到干预时的潜在结果。但任何人都只能观测到其中一种结果。实际中受到干预的人无法观测未受到干预时的潜在结果。而实际未受干预的人也无法观测到受到干预时的潜在结果。也就是说，鲁宾认为因果推理的根本问题就在于任何人都只能观测到一种潜在结果。

补论1　了解分析的有效性和局限性

内部效度与外部效度

以上我们学习了因果推理的多种方法。在这个补论中，我们将介绍一下这些方法的“有效性”与“局限性”。先来解释两个用来评价分析结果有效性的概念：“内部效度”和“外部效度”。

所谓内部效度，是指两个变量之间的因果关系的确定程度，即对研究对象群体再次施加相同干预后，会出现相同结果的程度；而外部效度则指对研究对象之外的群体施加相同干预后，会出现相同结果的程度。

我们来思考下面这个虚构的事例。假设美国进行的一项随机对照试验证实了以下因果关系：降低血压可以使人们患心肌梗死的风险下降，死亡率也随之下降。随机对照试验的证据等级较高，因此可认为该试验的内部效度也比较高。不过如果这项随机对照试验的对象是美国人，那么对其结果是否适用于日本人，即对该试验的外部效度则需要慎重考量。

通常认为，日本人患脑梗死的概率大于心肌梗死，而美国人患心肌梗死的概率大于脑梗死。为此，血压降低是否能够使日本人患心肌梗死的概率下降，这个课题必须用日本人的数据进行论证。只有当我

们针对日本人实施相同研究，并观测到降低血压会使人们患心肌梗死的概率下降的结果时，才能说该因果关系“具有外部效度”。

随机对照试验也有局限性

正如前面提到的，本书介绍的各种方法虽然都是学术上成立的方法，但并非完美无缺。我们只有了解每种方法的局限性，对所得结果的有效性进行充分的探讨，才能得出“两个变量之间是否存在因果关系”这一问题的准确答案。

本书虽然也介绍了多种利用自然实验或准实验根据观测数据判断因果关系的方法，不过就判断因果关系这一点而言，随机对照试验毫无疑问要优于利用观测数据进行的研究。不过，随机对照试验也有以下几点局限性。

首先，实施随机对照试验的成本很高。前文提到“兰德健康保险实验”投入了3亿美元的巨额费用，如此昂贵的社会实验不是任何人都承担得起的。

其次，存在“外部效度”的问题。很多随机对照试验在挑选研究对象时都设有严格的筛选条件，因此无法确定该研究结果是否也适用于其他群体。

第三，很多试验出于伦理方面的考虑无法实施。比如，研究者不能因为想研究吸烟与肺癌的因果关系，就强迫受试者吸烟。

第四，随机对照试验有时无法按计划“随机分组”，导致“随机化失败”。除了分组时会发生这种情况之外，在试验进行过程中，试验对象从对照组转移到干预组也会破坏随机性。

第五，与随机对照试验中得到的效果（efficacy）相比，实际在整个社会实践后产生的实效（effectiveness）可能会变小（该概念与前面

介绍的外部效度比较接近，不过鉴于这种方式或许更容易理解，因此分开介绍）。例如，在评估药效的临床实验中，实验对象都是经过严格挑选的。老年人或同时患有多种疾病的人往往被排除在受试者范围之外。然而，如果临床试验证明药物有效，老年人和同时患有多种疾病的人也会使用这种药物，因此实际上的药效会有别于临床试验得出的效果。

综上所述，随机对照试验并非绝对正确，它也有自身的局限性。随机对照试验的证据等级之所以很高，只是因为这种方法可以确保满足第1章介绍的判断因果关系的三个要点：（1）不是纯属巧合；（2）不存在混杂因素；（3）不存在逆向因果关系。反过来说，只要能在充分探讨的基础上确保满足以上三点，利用观测数据的研究也能成为有力证据。

补论2　因果推理的五个步骤

本节将因果推理总结为五个步骤（图表5），借此对本书介绍的内容进行系统性的复习。我们继续用广告和销售额的事例来解说。

图表5 解读因果关系的五个步骤

1.“原因”是什么

2.“结果”是什么

3.确认三个要点（参考正文第4页）

4.制造反事实

5.调整到可比较的状态

步骤1 “原因”是什么

指原因与结果中的“原因”，相当于广告与销售额事例中的“广告”。不过，广告也包括很多内容，比如是指具体的广告费用、投放面积，还是只是指有无投放广告？必须对原因做出明确的定义。

步骤2 “结果”是什么

指原因与结果中的“结果”，相当于广告与销售额事例中的“销售额”。想统计的是销售收入还是税前利润？结果也需要一个明确的定义。

步骤3 确认三个要点

前文提到过，判断因果关系时需要质疑三个要点：（1）是不是纯属巧合；（2）是否存在混杂因素；（3）是否存在逆向因果关系。

在广告与销售额的事例中，需要仔细分析是否存在能够同时影响广告和销售额的混杂因素。

比如，很可能存在“经济形势较好”等混杂因素。经济形势好的话，投放广告的决策更容易通过，销售额也更容易上涨。如果存在“经济形势较好”的混杂因素，广告与销售额的关系则可能只是伪相关。

步骤4 制造反事实

为了判断广告与销售额之间是否存在因果关系，你的公司需要对投放广告后的销售额和假设未曾投放广告时的销售额进行比较。这里的“假设未曾投放广告”就是“反事实”。例如，如果经济形势好，便可以想象“即使当初没有投放广告，销售额也会增加”，所以广告对销售额的实际效果或许会比我们预想的要小一些。

可是，我们无法得知“假设未曾投放广告”这种反事实中的销售额（毕竟没有时光机），而要评估广告与销售额的因果关系，反事实又是必不可少的要素，因此需要用“最贴切的值”来替换反事实的结果。也就是需要探索其他的可能性，比如能否使用在同一时期未投放广告的竞争对手的销售额数据等。

步骤5 调整到可比较的状态

用最贴切的值替换反事实的结果的一种有效方法是，通过调整形成“可比较”的组。例如假设在同一时期未投放广告的竞争对手和你的公司是“可比较”的，那么就可以用竞争对手的销售额作为最贴切的值，来替换“假设你的公司未曾投放广告（反事实）时的销售额”。

后　记

2016年1月8日的《华尔街日报》刊登了一篇题为“美国监管机构提示，警惕大数据中潜在的偏差”的报道。其内容是美国联邦贸易委员会向运用大数据分析商业活动的企业发出警告，具有相关关系并不代表存在因果关系。

相信各位读者都能理解这个警告的含义。在美国联邦贸易委员会列举的事例中，企业采用与经济状况或还款记录几乎没有任何关系的变量，对信用风险进行回归分析，预测个人信用风险的高低，并据此决定是否开展业务。这种方法得出的结果很可能只是本书中介绍的“纯属巧合”导致的伪相关（参考正文第5页），然而人们却有可能因为这些结果被剥夺了应得的机会。美国联邦贸易委员会对此敲响了警钟。

在“大数据”成为流行词的当下，数据分析随处可见。然而，数据本身只不过是罗列在一起的数字。“如何解读”数据分析才是关键。误将只呈现出相关关系的数据分析当作因果关系，会导致人们进一步做出错误的判断。

从这个意义上讲，“因果推理”如今已经不再只是笔者等研究人员的“专利”，而是所有人为了弄清楚“是否存在因果关系”所必备的素养。

在国外，基于因果推理得出的研究成果正在切实地改变着人们的生活。墨西哥第54任总统埃内斯托·塞迪略曾经开展过一项名为“进步计划”（PROGRESA）的扶贫项目，同时通过大规模随机对照试验，对进步计划的效果进行了严格测定。

那么，这样做的目的是什么呢？在每六年一次的总统换届选举之后，扶贫政策的方针都会发生巨大转变。塞迪略总统此举正是为了改变现状，使扶贫政策建立在具有因果关系的证据基础之上，而不再受各时期总统或政党意识形态的左右。如果随机对照试验能够明确证实“进步计划具有减少贫困的效果”，那么今后即使换作其他总统或政党执政，进步计划仍然会得到纳税人的支持，继续开展下去（事实上塞迪略总统离任之后，“进步计划”至今仍在继续实施）。

美国也有一个致力于依据证据牵制政策的组织。埃斯特·迪弗洛（Esther Duflo）等新锐经济学家组成的麻省理工学院扶贫实验室是一家专门开展随机对照试验的研究机构，为倡导“将容易受政治风向左右的政策变为建立在证据基础上的政策”而设立。他们的一系列研究正在显著改变着曾经被无稽之谈和传统观念影响的教育、医疗领域的政策制定过程。

本书作者之一的津川亲身经历过这样的情况。在他参加2014年10月哈佛大学召开的奥巴马医改计划研讨会时，台下一名记者发言称：“奥巴马医改实施后保险费用不降反升，我的好几个熟人都为此深感为难。奥巴马医改计划反而把美国的医疗制度改得更糟了。”对此，参与设计奥巴马医改计划的麻省理工学院医疗经济学家乔纳森・格鲁伯（Jonathan Gruber）答道：“个人经历的集合既不是数据，也称不上证据。我们通过缜密的数据收集验证了奥巴马医改计划的效果。其结果显示，美国国民的平均保险费用由于奥巴马医改计划的实施而有所下降。也许有些人因为不得不支付更多保险费而蒙受了损失，但是不要被个别现象误导，请用数据来呈现整体的变化。”

医疗费用关系到生命，正因如此，相关话题的讨论容易以个人经历为依据，带上个人情绪。格鲁伯等人的言论提倡不能感情用事，并

指出了以数据和证据为依据进行讨论的重要性。

再来看看日本。本书也多次提及在国会所在地永田町霞关针对各种政策展开的讨论，不过遗憾的是，目前政策方面的讨论还很难被视为以具有因果关系的证据为依据。不仅如此，每逢大选临近，政客们便只顾讨论能在短期内吸引选票的政策，他们常常违背之前的承诺，或者突然强行推行没有任何根据的政策，致使纳税人的利益蒙受严重损失。这样的事例不胜枚举。这正是不重视国民的未来、只顾眼前的选举所招致的结果。

每当目睹这种局面，我都会发出感慨：为什么不能摆脱选举、政局等暂时性政治作秀的影响，从长远的角度出发，来考虑如何实现国民的社会福利（幸福）最大化呢？为此，应该像扶贫实验室所宣称的，“把容易受到政治风向左右的政策变为建立在证据基础上的政策”。要实现这个目标，我们每一名国民作为纳税人也必须从“什么政策才具有因果效应”的角度出发，严格审视每项政策。

最后，请允许我向在本书出版过程中给予我莫大支持的人们表示衷心感谢。

实际上，本书从执笔构思至今已经过了两年的时间。学者著作的后记中，常会看到诸如“自开始写书到出版历经了四分之一个世纪”“献给如今已故的编辑〇〇先生”等表述，让人感慨光阴荏苒。本书也几乎踏上同样漫长而艰难的道路，好在在编辑的催促与鼓励之下，终于得以付梓。在此对负责本书的钻石社编辑上村晃大先生致以诚挚的谢意。

此外，统计学家兼因果推理专家、庆应义塾大学星野崇宏教授自本书草稿阶段便认真协助校对，并为我提出了很多宝贵建议。东京大学的朝井友纪子助教、一桥大学的神林龙教授、麦克马斯特大学的山口慎太郎副教授、西蒙弗雷泽大学的重冈仁副教授（以上人名按刊登

章节顺序排列）欣然同意本书引用并介绍其研究。在此对各位表示感谢。

书中介绍的笔者们的研究属于日本经济产业研究所（RIETI）的“对医疗及教育质量的测量及其决定因素的分析”（研究负责人：学习院大学乾友彦教授）和“劳动市场制度改革”（研究负责人：庆应义塾大学鹤光太郎教授）项目中的部分内容。以矢野诚所长和森川正之副所长为首的经济产业研究所为我们提供了大力支持，在此表示感谢。另外，科研经费补助基础S“社会障碍的经济理论与实证研究”（研究负责人：东京大学松井彰彦教授）、基础B“关于幼年期所处社会生活环境和学习方法对人力资本积累的影响的分析”（研究负责人：信息安全大学院大学广松毅教授）、基础A“贫困与灾害的教育经济学：如何培养出能够战胜社会不利因素和困难的孩子”（研究负责人：中室牧子）也为本书的编写提供了支援。

感谢庆应义塾大学综合政策学院政策媒体研究科中室研究室的植村理、川崎美波、坂本彩乃、中川舞音、中田知宏、中村真优子、锅泽步、山越梨沙子、吉屋麻里为本书的草稿提出了宝贵建议。

“研究”是一份奇特的工作。无论撰写多少篇论文，在研讨会上做多少场演讲，我们也不会变得更加富有。不过，研究中诞生的观点可以作为公共知识财富，为建设更美好的世界做出贡献。感谢我们的家人对这份工作给予的理解和毫无保留的支持。特别是在本书草稿阶段多次过目并提出建议的代田丰一郎和津川衣林梨，在此向你们致以由衷的感谢。最后需要声明的是，本书虽然是在众多人士的支持下完成的，但书中任何错误均由笔者负责。

“在证据的基础上制定政策”的做法在国外已经扎根，但在教育和医疗领域，无稽之谈和传统观念的影响力依旧很强，想在证据的基础上制定政策绝非易事。为了实现这个目标，我们这些研究者也许要倾尽一生的时间，不断推广这些观点才行。

证据能够体现经济学所信奉的“因果关系”，只有通过体系化的“因果推理”才能形成证据。如果这些工具能够对生活在数据泛滥时代的各位读者有所帮助，我们身为作者也会感到欣慰。

2017年2月
中室牧子
津川友介

参考文献

第1章

タイラー・ヴィーゲン ウェブサイト（http://tylervigen.com/spurious-correlations）

Vigen, T. (2015) *Spurious Correlations Hardcover*, Hachette Books.

Messerli, F. H. (2012) Chocolate Consumption, Cognitive Function, and Nobel Laureates, *The New England Journal of Medicine*, 367, 1562-1564.

Brickman, A. M., Khan, U. A., Provenzano, F. A., Yeung, L., Suzuki, W., Schroeter, H., Wall, M., Sloan, R. P. and Small, S. A. (2014) Enhancing Dentate Gyrus Function with Dietary Flavanols Improves Cognition in Older Adults, *Nature Neuroscience*, 17 (12), 1798-1803.

Sackett, D. L., Straus, S. E., Richardson, W. S., Rosenberg, W. and Haynes, R. B. (2000). *Evidence-based Medicine: How to Practice and Teach EBM*, Churchill Livingstone, 2 edition.

第2章

Krogsbøll, L. T., Jørgensen, K. J., Grønhøj Larsen, C. and Gøtzsche, P. C. (2012) General Health Checks in Adults for Reducing Morbidity and Mortality from Disease: Cochrane Systematic Review and Meta-analysis, *BMJ*, 345, e7191.

Jørgensen, T., Jacobsen, R. K., Toft, U., Aadahl, M., Glümer, C. and Pisinger, C. (2014) Effect of Screening and Lifestyle Counselling on Incidence of Ischaemic Heart Disease in General Population: Inter99 Randomised Trial, *BMJ*, 348, g3617.

「メタボ健診 システム不備 効果検証、2 割しかできず 会計検査院、改修求める」『日本経済新聞』2015年9月5日朝刊、38頁

「日本の医療費は高額 新基準で世界 3 位―対 GDP、OECD まとめ」『日本 経済新聞電子版ニュース』2016 年 8 月 21 日（http://www.nikkei.com/article/DGXLASFS18H1I_Y6A810C1SHA000/）

Manning, W. G., Newhouse, J. P., Duan, N., Keeler, E. B., Leibowitz, A. and Marquis, M. S. (1987) Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment, *American Economic Review*, 77 (3), 251-277.

「がんセンターとJT、肺がんリスク巡り対立 疑義に反論」朝日新聞デジタル 2016 年 9 月 30 日（http://www.asahi.com/articles/ASJ9Y5GC8J9YULBJ00J.html）

「受動喫煙による日本人の肺がんリスク約1.3倍 肺がんリスク評価『ほぼ確実』から『確実』へ」国立がんウェブサイト（http://www.ncc.go.jp/jp/information/press_release_20160831.html）

「受動喫煙と肺がんに関わる国立がん研究センター発表に対する JT コメント」 JT ウェブサイト（ https:// www. jt i.co. jp/ tobacco/ responsibilities/opinion/ f sc_ report/20160831.html）

「受動喫煙と肺がんに関する JT コメントへの見解」国立がん研究センターウェブサイト（http://www.ncc.go.jp/jp/information/20160928.html）

Hori, M., Tanaka, H., Wakai, K., Sasazuki, S. and Katanoda, K. (2016) Secondhand Smoke Exposure and Risk of Lung Cancer in Japan: A Systematic Review and Meta-Analysis of Epidemiologic Studies, *Japanese Journal of Clinical Oncology*, 46 (10), 942-951.

第3章

Tsugawa, Y., Jena, A. B., Figueroa, J. F., Orav, E. J., Blumenthal, D. M. and Jha, A. K. (2017) Comparison of Hospital Mortality and Readmission Rates for Medicare Patients Treated by Male vs Female Physicians, *JAMA Internal Medicine*, 177 (2), 1-8.

Royer, H. (2009) Separated at Girth: US Twin Estimates of the Effects of Birth Weight, *American Economic Journal: Applied Economics*, 1 (1), 49-85.

Black, S. E., Devereux, P. J. and Salvanes, K. G. (2007) From the Cradle to the Labor Market? The Effect of Birth Weight on Adult Outcomes, *The Quarterly Journal of Economics*, 122 (1), 409-439.

Oreopoulos, P., Stabile, M., Walld, R. and Roos, L. L. (2008) Short-, Medium-, and Long-Term Consequences of Poor Infant Health: An Analysis Using Siblings and Twins, *The Journal of Human Resources*, 43 (1), 88-138.

Lin, M. J., and Liu, J. T. (2009) Do Lower Birth Weight Babies Have Lower Grades? Twin Fixed Effect and Instrumental Variable Method Evidence from Taiwan, *Social Science & Medicine*, 68 (10), 1780-1787.

Nakamuro, M., Uzuki, Y., and Inui, T. (2013) The Effects of Birth Weight: Does Fetal Origin Really Matter for Long-run Outcomes?. *Economics Letters*, 121 (1), 53-58.

Paul, A. M. (2011) Origins: How the Nine Months Before Birth Shape the Rest of Our Lives, Free Press.

Ferrante, D., Linetzky, B., Virgolini, M., Schoj, V. and Apelberg, B. (2012) Reduction in Hospital Admissions for Acute Coronary Syndrome after the Successful Implementation of 100% Smoke-free Legislation in Argentina: A Comparison with Partial Smoking Restrictions. *Tobacco Control*, 21 (4), 402-406.

González-Rozada, M., Molinari, M. and Virgolini, M. (2008) The Economic Impact of Smoke-free Laws on Sales in Bars and Restaurants in Argentina, *CVD Prevention and Control*, 3 (4), 197-203.

Glantz, S. A. and Smith, L. R. (1997) The Effect of Ordinances Requiring Smoke-free Restaurants and Bars on Revenues: A Follow-up, *American Journal of Public Health October*, 87 (10), 1687-1693.

Glantz, S. A. and Charlesworth, A. (1999) Tourism and Hotel Revenues before and after Passage of Smoke-free Restaurant Ordinances, *JAMA*, 281 (20), 1911-1918.

第4章

Asai, Y., Kambayashi, R. and Yamaguchi, S. (2015) Childcare Availability, Household Structure, and Maternal Employment, *Journal of the Japanese and International Economies*, 38, 172-192.

山口慎太郎（2016）「差の作法で検証する「保育所整備」の効果」、岩波データサイエンス刊行委員会編 『岩波データサイエンス Vol.3 ［特集］因果 推論—実世界のデータから因果を読む』岩波書店.

朝井有紀子・神林龍・山口慎太郎（2016）「保育所整備と母親の就業率」『経 済分析』第191号（特別編集号），123-152.

Heckman, J. J. (2006) Skill Formation and the Economics of Investing in Disadvantaged Children. *Science*, 312 (5782), 1900-1902.

Card, D. and Krueger, A. B. (1994) Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania, *American Economic Review*, 84 (4), 772-793.

Card, D. and Krueger, A. B. (2000) Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Reply, *American Economic Review*, 90 (5), 1397-1420.

Dube, A., Lester, T. W. and Reich, M. (2010) Minimum Wage Effects Across State Borders: Estimates Using Contiguous Counties, *The Review of Economics and Statistics*, 92 (4), 945-964.

大竹文雄・川口大司・鶴光太郎編著（2013）『最低賃金改革：日本の働き方をいかに変えるか』日本評論社

鶴光太郎（2013）「最低賃金の労働市場・経済への影響—諸外国の研究から得られる鳥瞰図的な視点—」*RIETI Discussion Paper Series*, 13-J-008.

Neumark, D. and Wascher, W. (2000) Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Comment, *American Economic Review*, 90 (5), 1362-1396.

Straight, S. (2003) Other Juvenile Awareness Programs for Preventing Juvenile Delinquency: A Systematic Review of the Randomized Experimental Evidence, *Annals of the American Academy of Political and Social Science*, 589, 41-62.

Farrington, D. P., and Welsh, B. C. (2005) Randomized Experiments in Criminology: What Have We Learned in the Last Two Decades?. *Journal of Experimental Criminology*, 1 (1), 9-38.

第5章

Gentzkow, M. and Shapiro, J. M. (2008) Preschool Television Viewing and Adolescent Test Scores: Historical Evidence from the Coleman Study, *The Quarterly Journal of Economics*, 123 (1), 279-323.

Currie, J. (2009) Healthy, Wealthy, and Wise: Socioeconomic Status, Poor Health in Childhood, and Human Capital Development, *Journal of Economic Literatur*, *e*47 (1), 87-122.

Currie, J. and Almond, D. (2011) Human Capital Development before Age Five, *Handbook of Labor Economic*, *s*4, 1315-1486.

Currie, J. and Moretti, E. (2003) Mother's Education and the Intergenerational Transmission of Human Capital: Evidence from College Openings, *The Quarterly Journal of Economics*, 1495-1532.

Greenstone, M. and Looney, A. (2011) Where Is the Best Place to Invest $102, 000: In Stocks, Bonds, or a College Degree?, *Hamilton Project*.

Ahern, K.R. and Dittmar, A. K. (2012) The Changing of the Boards: The Impact on Firm Valuation of Mandated Female Board Representation, *The Quarterly Journal of Economics*, 127 (1), 137-197.

Joshi, A. and Roh, H. (2009) The Role of Context in Work Team Diversity Research: A Metaanalytic Review, *Academy of Management Journal*, 52 (3), 599-627.

Østergaard, C. R., Timmermans, B. and Kristinsson, K. (2011) Does a Different View Create Something New? The Effect of Employee Diversity on Innovation, *Research Policy*, 40 (3), 500-509.

第6章

Abdulkadiroğ lu, A., Angrist, J. and Pathak, P. (2014) The Elite Illusion: Achievement Effects at Boston and New York Exam Schools, *Econometrica*, 82 (1), 137-196.

Kling, J. R., Liebman, J. B., and Katz, L. F. (2007) Experimental Analysis of Neighborhood Effects, *Econometrica*, 75 (1), 83-119.

Shigeoka, H. (2014) The Effect of Patient Cost Sharing on Utilization, Health, and Risk Protection, *American Economic Review*, 104 (7), 2152- 2184.

照山博司・細野薫・松島斉・松村敏弘編（2016）「医療・介護の持続可能性と経済学：パネル討論I」，『現代経済学の潮流2016』東洋経済，165-200.

重岡仁「気鋭の論点：医療費の「高齢者1割負担」がもたらすメリットとデメリット」2014年2月20日（木）日経ビジネスオンライン

Rossouw, J. E., Anderson, G. L., Prentice, R. L., LaCroix, A. Z., Kooperberg, C., Stefanick, M. L., Jackson, R. D., Beresford, S. A., Howard, B. V., Johnson, K. C., Kotchen, J. M. and Ockene, J.; Writing Group for the Women's Health Initiative Investigators (2002) Risks and Benefits of Estrogen plus Progestin in Healthy Postmenopausal Women: Principal Results from the Women's Health Initiative Randomized Controlled Trial, *JAMA*, 288(3), 321-333.

Manson, J. E., Hsia, J., Johnson, K. C., Rossouw, J. E., Assaf, A. R., Lasser, N L., Trevisan, M., Black, H. R., Heckbert, S. R., Detrano, R., Strickland, O L., Wong, N. D., Crouse, J. R., Stein, E. and Cushman,

M.; Women's Health Initiative Investigators (2003) Estrogen plus Progestin and the Risk of Coronary Heart Disease, *The New England Journal of Medicine*, 349 (6), 523-534.

第7章

Dale, S. B. and Krueger, A. B. (2002) Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables, *The Quarterly Journal of Economics*, 117 (4), 1491-1527.

Dale, S. B. and Krueger, A. B. (2014) Estimating the Effects of College Characteristics over the Career Using Administrative Earnings Data, *The Journal of Human Resources*, 49 (2), 323-358.

Kambayashi, R., Kawaguchi, D. and Yokoyama, I. (2008) Wage Distribution in Japan, 1989-2003, *Canadian Journal of Economics*, 41 (4), 1329-1350.

Black, D. A., and Smith, J. A. (2004) How Robust Is the Evidence on the Effects of College Quality? Evidence from Matching, *Journal of Econometrics*, 121 (1-2), 99-124.

Bertrand, M., Karlan, D., Mullainathan, S., Shafir, E., and Zinman, J. (2010) What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment, *The Quarterly Journal of Economics*, 125 (1), 263-306.

第8章

Korte, J. E., Brennan, P., Henley, S. J. and Boffetta, P. (2002) Dose-Specific Meta-Analysis and Sensitivity Analysis of the Relation between Alcohol Consumption and Lung Cancer Risk, *American Journal of Epidemiology*, 155 (6), 496-506.

Freudenheim, J. L., Ritz, J., Smith-Warner, S. A., Albanes, D., Bandera, E. V., van den Brandt, P. A., Colditz, G., Feskanich, D., Goldbohm, R. A., Harnack, L., Miller, A. B., Rimm, E., Rohan, T. E., Sellers, T. A.,

Virtamo, J., Willett, W. C. and Hunter, D. J. (2005) Alcohol Consumption and Risk of Lung Cancer: A Pooled Analysis of Cohort Studies, *The American Journal of Clinical Nutrition*, 82 (3) 657-667.

Bagnardi, V., Rota, M., Botteri, E., Tramacere, I., Islami, F., Fedirko, V., Scotti, L., Jenab, M., Turati, F., Pasquali, E., Pelucchi, C., Galeone, C., Bellocco, R., Negri, E., Corrao, G., Boffetta, P. and La Vecchia, C. (2015) Alcohol Consumption and Site-Specific Cancer Risk: A Comprehensive Dose-Response Meta-Analysis, *British Journal of Cancer*, 112 (3), 580-593.

Shimazu, T., Inoue, M., Sasazuki, S., Iwasaki, M., Kurahashi, N., Yamaji, T. and Tsugane, S.: Japan Public Health Center-based Prospective Study Group. (2008) Alcohol and Risk of Lung Cancer among Japanese Men: Data from a Large-Scale Population-Based Cohort Study, the JPHC Study, *Cancer Causes and Control*, 19 (10), 1095-1102.

補論

Imbens, G. W. and Rubin, D. B. (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press.

Pearl, J. (1995) Causal Diagrams for Empirical Research, *Biometrika*, 82 (4), 669-688.

后浪

简单统计学

Standard Deviations

Flawed Assumptions, Tortured Data, and Other Ways to Lie with Statistics

[美] 加里·史密斯／著
Gary Smith
刘清山／译

如何轻松识破一本正经的胡说八道

让数据说实话，是现代社会的基本生存技能

章鱼保罗、美职篮、房市泡沫、畅销书、投资……

只需懂加减乘除，耶鲁大学统计学公开课
帮你拆穿日常生活中的数字谎言

诺贝尔经济学奖得主
罗伯特·希勒 | 赞赏推荐

江西人民出版社

简单统计学

[美]加里·史密斯 著
刘清山 译

版权信息

书名：简单统计学
作者：[美]加里·史密斯
译者：刘清山
书号：ISBN 978-7-210-09841-6
版权：后浪出版咨询（北京）有限责任公司

目录

序言

第 1 章　模式、模式、模式

- 混杂效应
- 选择性报告与谎报
- 易受欺骗的本性
- 无论文，不生存
- 统计显著性膜拜

第 2 章　不再神奇的超级畅销书

- 去最好的学校
- 投票人数越多越好？
- 一醉方休
- 放下遥控器
- 请原谅我的直白
- 只有弱者留下来
- 被损坏的飞机
- 畅销书的秘密

第 3 章　被误传的谋杀之都

马萨诸塞州谋杀之都

请在我家后院开一座采石场

第 4 章　新的经济学上帝

政府债务临界点

相关性等同于因果关系？

堕胎会减少犯罪吗？

是我的错

第 5 章　扬基队的门票真的划算吗？

我的天哪

虚惊一场

让他们吃蛋糕吧

有弹性的数轴

将恶作剧翻倍

学生谎言

扬基队的门票很划算

图像的制作艺术

第 6 章　美国有多少非畜职业运动员？

蒙提·霍尔问题

一个名叫佛罗里达的女孩儿

条件概率的混淆

假阳性问题
罕见病问题
达特茅斯三文鱼研究
第 7 章　辛普森悖论
某人的悖论
我要再来一杯咖啡
第 8 章　状态火热的雷·阿伦
小数定律
一项篮球研究
小沃尔特·雷·威廉姆斯
投掷马蹄铁
保龄球
第 9 章　胜者的诅咒
均值回归
西克里斯特的愚蠢
古老的谬误何曾消亡？
道指落榜生
冠军窒息
寻找学院院长和灵魂伴侣
第 10 章　如何转变运气？
第 11 章　德克萨斯神枪手

癌症聚集恐慌

第 12 章　终极拖延

死亡凹陷和尖峰

经过进一步研究

午夜的月饼

漫长的告别

第 13 章　黑色星期一

火箭科学

巴斯克维尔猎犬

五行噩运

花押决定论

第 14 章　点球成金

棒球迷信

如果你给我一个 D，我就会死（die）

糟糕的出生月份

名人堂的死亡之吻

第 15 章　特异功能真的存在吗？

超感知觉

霍迪尼的挑战

一个家庭对超自然现象的着迷

第 16 章　彩票是一种智商税

目之所及
我们都将为 IBM 工作
股票是一种不错的投资
跑赢大盘（或者说为什么我喜爱包装工队）
傻瓜四股
反向头肩
如何（不）中彩票
不会破裂的泡沫
南海泡沫
伯克希尔泡沫
真实股价
第 17 章　超级投资者
宽客
收敛交易
我只看数据
在推土机面前捡硬币
闪电崩盘
第 18 章　增长的极限
穷途末路
第 19 章　何时相信，何时怀疑
被模式诱惑
具有误导性的数据

变形的图像
缺乏思考的计算
寻找混杂因素
手气好
均值回归
平均定律
德克萨斯神枪手
当心经过修剪的数据
缺乏理论的数据仅仅是数据而已
缺乏数据的理论仅仅是理论而已
美好的出生日
出版后记
返回总目录

序言

我们生活在大数据时代。高性能计算机和全球网络的强大组合正在得到人们的赞美甚至推崇。专家不断告诉我们，他们发现了一种新的能力，可以对海量数据进行筛查并发现真相，这将为政府、商业、金融、医疗、法律以及我们的日常生活带来一场革命。我们可以做出更明智的决策，因为强大的计算机可以对数据进行分析，发现重要的结论。

也许事实的确如此，也许未必。有时，这些无所不在的数据和伟大光明正确的计算机会得出一些非常怪异的结论。例如，有人一本正经地宣称：

- 凌乱的房间会强化人们的种族主义倾向。
- 还未出生的小鸡胚胎会对计算机的随机事件生成器产生影响。
- 当政府负债相对国内生产总值的比率超过 90% 时，国家几乎一定会陷入衰退。
- 在过去 20 年美国犯罪率下降的原因中，合法堕胎的比例高达 50%。

- 如果每天饮用两杯咖啡，患上胰腺癌的风险将极度放大。
- 最成功的公司倾向于变得不那么成功，最不成功的公司倾向于变得更加成功，因此用不了多久，所有公司都会沦为普通的公司。
- 出现在《体育画报》和《麦登橄榄球》封面上的运动员会受到诅咒，他们可能会陷入平庸，或者受到伤病困扰。
- 生活在输电线附近的儿童具有更大的患癌风险。
- 人类有能力将死亡推迟到重大仪式过后。
- 亚裔美国人更容易在每月四号突发心脏病。
- 如果一个人的姓名首字母缩写拥有积极的含义（比如 ACE），那么他可以多活三到五年。
- 平均来说，教名（第一个名字）以字母 D 开头的棒球运动员的寿命比教名以字母 E 到 Z 开头的运动员短两年。
- 临终病人可以被几千英里以外传送过来的积极心理能量治愈。
- 当 NFC 冠军球队赢得超级碗（美国超级碗总决赛在国家联合会（NFC）和美国联合会（AFL）分别比赛产生的冠军之间举行——译者注）时，股市几乎一定会上涨。
- 如果你购买股息率最高、每股价格第二低的道琼斯股票，你就可以跑赢大盘。

这些说法显然是错误的。不过，许多与此类似的说法每天都会出现在报纸和杂志上。在如今的信息时代，我们用没完没了而又毫无意义的数据指导我们的思想和行动。不难看出为什么我们会反复得出错误的推论，制定糟糕的决策。即使能够得到比较充足的信息，我们也不会永远注意到数据的偏差性和无关性，或者科学研究的缺陷和误导性。我们倾向于相信计算机从不犯错，认为不管我们把什么样的垃圾扔进去，计算机都会吐出绝对真理。这种想法不仅存在于外行人的日常生活中，也存在于专业人员严肃认真的研究工作中。在流行刊物、电视、互联网、竞选活动、学术期刊、商业会议、法庭，政府听证会，此类现象屡见不鲜。

几十年前，数据非常稀少、计算机还没有出现时，研究人员需要努力收集优质数据并进行审慎的思考，然后花费几个小时甚至几天的时间从事艰苦的计算工作。现在，面对丰富的数据，研究人员通常不会花费太多的时间对优质数据和垃圾进行区分，或者对合理分析和垃圾科学进行区分。更糟糕的是，我们常常不假思索地认为，我们对大量数据的处理永远不会出错。我们匆匆忙忙地根据这些机器发出的梦呓制定决策——比如在衰退过程中增加税收，将我们一生的积蓄交给一些说得天花乱坠的财务分析师，根据最新的管理理念制定商业决策，用医疗骗术危害我们的健康——更糟糕的是，我们还会放弃心爱的咖啡。

罗纳德·科斯（Ronald Coase）曾经嘲讽道：“如果你对数据拷打足够长的时间，它一定会招供。”《简单统计学》一书考察了几十个扭曲的结论。只需片刻的思考，你就会发现这些结论的问题。有时，无耻之徒故意用这些说法来误导我们。有时，天真快乐的研究人员并没有意识到他们所制造的恶作剧。我写这本书的目的是帮助我们远离错误——包括外部错误和自己造成的错误。你将学到一些简单的指导准则，用于识别其他人或者你自己说出的不靠谱的观点。其他人用数据欺骗我们，我们也经常用数据欺骗自己。

第 1 章
模式、模式、模式

在我所生活的南加州，青少年对于足球的热情很高。这是一项经济而有趣的运动，各种身高和体型的男生和女生都可以参与其中。起初，我对足球一无所知。我只知道在每个周末，城市里的公园和操场上都会有许多身穿鲜艳队服的孩子追着足球跑来跑去，他们的家长则会站在一旁为他们欢呼。当我的儿子长到合适的年龄时，我们也成了这些家长中的一员。

2010 年世界杯时，我的儿子在南加州一家顶级足球俱乐部踢球。我是这家俱乐部的经理，也是足球的狂热爱好者，因此我们共同观看了我们能够看到的每一场世界杯比赛。决赛是在荷兰和西班牙之间进行的。这是两支巨星云集的队伍，但是它们所在的国家常常无法取得很好的成绩，这使它们的支持者非常失望。哪个国家最终能够赢得世界杯？我喜爱荷兰队，他们赢得了这届世界杯之前的 6 场比赛，进了 12 个球，只丢了 5 个球，而且淘汰了强大的巴西和乌拉圭。接着，我听说了章鱼保罗（Paul）的故事。人们将装有食物、贴有国旗的塑料盒子放在保罗面前，让它进行选择。通过这种方式，保罗正确预测出了 7 场世界杯比赛的获胜者。在决赛前，预言家保罗选择了西班牙，因此全世界似乎都认为西班牙将会成为胜利者。

到底发生了什么事情？一只黏糊糊的、缺乏智商的无脊椎动物怎么可能比我更加了解足球呢？我感到好笑，等待着无所不知的保罗在

全世界面前丢脸。我失算了。荷兰队没有表现出应有的创造性和才华。在一场粗暴而蹩脚的比赛中，裁判亮出了 14 张黄牌——其中 9 张给了肮脏的荷兰人。凭借终场前 4 分钟的进球，西班牙队取得了胜利。

生活在水箱里的章鱼怎么可能预见到这种事情呢？保罗看过足球比赛吗？它有大脑吗？

实际上，章鱼是最聪明的无脊椎动物。不过，就像矬子里拔出来的将军一样，这并不能说明任何问题。尽管如此，保罗仍然进行了 8 次世界杯预测，而且无一失手。此外，保罗还在 2008 年欧洲足球锦标赛期间进行了 6 次预测，成功了 4 次。加上 2010 年世界杯，保罗一共进行了 14 次预测，成功了 12 次。在许多人看来，这足以从统计上证明保罗的超自然能力。不过，这些数据真的足够多吗？

如果进行 14 次公平的抛硬币实验，得到 12 次或者 12 次以上正面的可能性只有不到 1%。同样的道理，如果保罗仅仅是一个没有特殊运气的猜测者，每次正确预测的可能性为 50%，那么它多次预测成功的概率只有不到 1%，这个概率非常低，足以被视作具有“统计显著性”。由于保罗多次预测成功的可能性如此之低，因此我们可以合理地排除运气的成分。保罗用一次又一次的成功证明了它不仅仅是一个幸运的猜测者。它的确是“具有超能力的章鱼保罗”！

不过，这件事似乎有些蹊跷。章鱼真的有可能预测未来吗？保罗的表现可以引出统计研究领域的一些常见问题。保罗不是超能力者（真是令人惊喜），而是一个应当引以为戒的反面教材。当你下次听到某种荒诞的说法时，你应当保持警惕。

混杂效应

首先，让我们看一看保罗是如何进行预测的。在喂食的时候，人们会把两只正面粘有球队国旗的透明塑料盒子摆放在保罗面前。盒子里面放着相同的美味食物，比如贻贝或者牡蛎。保罗第一个打开的盒子就是它所预测的胜利者。

章鱼并不十分了解足球，但它们拥有出色的视力和良好的记忆力。有一次，新英格兰水族馆的一只章鱼对一名志愿者产生了反感。每当它看到这名志愿者时，它都会向她喷射海水。后来，这名志愿者离开了水族馆，上了一所大学。当她几个月以后回来时，章鱼仍然记得她，并且立即用海水淋湿了她的衣服。西雅图一家水族馆做了一项实验，一名志愿者为章鱼喂食，另一名穿着同样衣服的志愿者用木棍挑逗章鱼。一个星期以后，大多数章鱼都能将两个人区分开。当它们看到“好人”时，它们会向他靠近；当它们看到“坏人”时，它们会离开（有时还会向他喷射许多海水）。

“具有超能力的章鱼保罗”生活在德国的一家水族馆里。除了西班牙和荷兰的世界杯决赛，保罗只预测了德国队参加的比赛。在德国参加的 13 场比赛中，保罗 11 次选择了德国——而德国赢下了其中的 9 场比赛。保罗之所以选择德国，是因为它对德国的对手进行了详细的分析，还是因为它喜欢德国国旗？保罗几乎一定是色盲，不过实验表明，章鱼能够识别明暗度，而且喜欢横向形状。德国国旗有由三块鲜艳的水平条纹组成，塞尔维亚和西班牙的国旗也是如此，而保罗只选择过这三个国家。实际上，西班牙和德国的国旗非常相似，这也许可以解释为什么保罗在西班牙和德国之间的两次比赛中选了一次西班牙，并在世界杯决赛中选择了西班牙而不是荷兰。保罗只有一次没有选择德国或西班牙国旗，那是塞尔维亚和德国之间的一场比赛。

国旗显然是一个混杂因素，因为保罗选择的并不是最佳足球队，而是它最喜欢的国旗。说到底，“无所不知的保罗”只是一只缺乏智商的章鱼而已。



图 1.1　保罗最喜欢的国旗

选择性报告与谎报

对于保罗的成功，另一种解释是，许多人多次尝试过这种愚蠢的宠物把戏，用宠物来预测体育、彩票和股票领域的获胜者。

在 1000 个抛硬币的人之中，一定会有一些人连续抛出 10 次正面。同样的道理，在这些尝试宠物把戏的人之中，一定会有一些人取得成功。你觉得谁会得到报道呢？是选中获胜者的章鱼，还是无法做出成功预测的鸵鸟？

几年前，《达拉斯晨报》的一位体育专栏作家度过了极为倒霉的一个星期，他在选择国家橄榄球联盟（NFL）比赛的获胜球队时猜对了 1 次，猜错了 12 次，还有一场比赛是平局。他写道："理论上说，如果让达拉斯动物园里的一只狒狒观看 14 场 NFL 比赛的赛程，让它为每场比赛选择一支球队，它可以至少选出 7 支获胜球队。"第二个星期，拉达斯动物园里的大猩猩"坎达大帝"（Kanda the Great）通过选择训练员手里的纸条进行了预测。结果，坎达猜对了 9 次，猜错了 4 次，击败了《达拉斯晨报》的所有 6 位体育专栏作家。媒体像发现食物的饿狼一样争相报道这个故事。不过，如果坎达的表现没有这么好，比如猜对 6 次，猜错 7 次，它还会被人报道吗？

明尼苏达州苹果谷市明尼苏达动物园的官员也不甘示弱，他们表示，一只叫做明迪（Mindy）的海豚成功预测出了 NFL 比赛的结果。他们制作了一些树脂玻璃片，每个玻璃片上写有不同球队的名称。他们将每场比赛对应的两块树脂玻璃片扔到明迪的游泳池里，被明迪交还给训练员的那块玻璃片被视作它的“预测”。训练员表示，明迪在 53 场比赛中猜对了 32 场。根据这种说法，明迪的成功率为 60%，这足以帮助人们借此赢得橄榄球比赛的赌博而获利。

有多少鸟儿、蜜蜂和野兽曾经尝试预测 NFL 比赛、并且由于预测失败而没有得到报道呢？我们并不知道这一点，而这正是问题的关键。如果数百只宠物曾经被迫进行毫无意义的预测，而且我们只知道其中得到报道的成功宠物，没有考虑到数百只没有得到报道的失败宠物，我们就会产生错误的想法。

这个问题不限于橄榄球领域。明尼阿波利斯的一名股票经纪人曾经吹嘘说，当他选择股票时，他会把《华尔街日报》铺在地板上，让它的金毛寻回犬把右前爪放在报纸上，然后选择它的第一个指甲碰到的股票。他认为这种说法可以吸引投资者的关注，这一事实说明他本人可能是有问题的——他的顾客可能也有问题。

另一个因素是，为了出风头，人们可能会捏造数据，以吸引公众的注意。明尼阿波利斯那个股票经纪人和他的狗每天早上都会得到公正的监督吗？过去，桥牌曾经是美国最流行的扑克游戏，当时有许多人向当地报纸表示，他们抽到了同一花色的十三张牌。一位喜欢数学的桥牌选手通过估计发现，出现这种情况的数量太过离谱。考虑到抽到这种牌的可能性，当时进行过的总局数远远不足以产生如此众多的“同花色局”。值得注意的是，报纸报道的这种相同花色通常是黑桃。显然，为了让自己的名字出现在报纸上，人们对自己的经历进行了润色。

在章鱼保罗获得全世界的关注以后，新加坡一个之前默默无闻的算命先生表示，他的助手、长尾小鹦鹉马尼（Mani）正确预测出了世界杯四分之一决赛的所有四支获胜球队。于是，马尼获得了全世界的

关注。接着，它对后面的比赛进行了预测，认为在半决赛中乌拉圭将击败荷兰，西班牙将击败德国，在决赛中，西班牙将击败乌拉圭。在荷兰战胜乌拉圭以后，马尼改变了它对决赛的预测，选择了荷兰，这个结果当然又错了。不过，前来拜访这位算命先生的顾客从每天 10 个人增加到了每小时 10 个人——面对这个结果，你可能会产生疑问：这位主人的动机真的仅仅是娱乐吗？他最初提供的马尼对四分之一决赛的预测结果是真的吗？

保罗和马尼本应默默无闻，处于无人知晓的状态。为什么它们会声名鹊起、受到足球爱好者的认真对待，甚至受到他们的赞美和诅咒？真正有问题的不是它们，而是我们自己。

易受欺骗的本性

一个多世纪以前，夏洛克·福尔摩斯（Sherlock Holmes）向他饱受折磨的朋友华生（Watson）恳求道："数据！数据！数据！我不能在没有黏土的情况下制造砖块。"今天，福尔摩斯的愿望已经得到了满足。强大的计算机可以对海量数据进行筛选。问题不再是我们没有足够的数据，而是我们眼前的数据对我们产生了误导。这不完全是我们的错。我们可以将责任归咎于我们的祖先。

某些性状的演化原理比较简单。如果生物的某些可遗传特点能够帮助它们生存和繁衍，那么同那些在其他方面相似，但是没有这些特点的个体相比，拥有这些特点的个体更容易将它们传给未来的后代。经过一代又一代的传递，这些可遗传的特点将成为群体的主流。

桦尺蠖著名的演化历程是一个简单而直观的例子。这种飞蛾大部分时间生活在树上，它们通常是浅色的，这种颜色可以帮助它们躲避鸟儿的捕食。1848 年，英国发现了第一只深色桦尺蠖。到了 1895 年，曼彻斯特 98% 的桦尺蠖都是深色的。到了 20 世纪 50 年代，情况

再次出现了逆转。目前。深色飞蛾已经很少见了，而且可能很快就会灭绝。

进化论的解释是，深色飞蛾的出现与工业革命导致的污染相重合。煤灰和烟雾导致树木变黑，这使深色飞蛾获得了优势，因为它们可以更好地伪装自己，躲过捕食者的视线。由于深色飞蛾存活下来并繁衍后代的可能性得到了提高，因此它们开始成为基因池的统治者。后来，英国的空气清洁法律扭转了这一局面，因为浅色飞蛾可以更好地在没有污染的树上伪装自己。它们的生存优势使它们再次焕发了生机。

自然选择的其他例子更加微妙。例如，许多研究发现，男性和女性更喜欢具有对称面孔和身体的人。这不仅仅是一种文化因素——它适用于不同的社会，适用于婴儿，甚至适用于其他动物。在一项实验中，研究人员剪掉了一些雄性家燕的尾羽，使它们失去了对称性。其他雄性家燕则保留了对称的尾羽。当他们在这个基因池中放开雌性家燕时，这些雌性家燕选择了带有对称羽毛的雄性家燕。这种对于对称性的偏爱并不仅仅是一种肤浅的行为。缺乏对称性的潜在配偶显然存在某种基因缺陷，可能影响它的力量、健康和生育力。拥有对称性偏好的个体最终将成为基因池的统治者，而缺乏这种偏好的个体拥有强壮、健康和可育后代的可能性要低一些。

进化也是许多人对保罗和马尼信以为真的原因，尽管你可能不相信这一点。我们对于对称性的内在偏好，可以很好地说明我们的人类祖先是如何通过识别各种模式在这个残酷的世界上生存和繁衍的。乌云常常是降雨的前奏。灌木丛中的声音可能是捕食者发出的。毛发质量是繁殖力的象征。一些远古祖先可以通过模式识别更好地寻找食物和水源，发现危险，对可育伴侣产生吸引力，他们将这种能力传给了未来的后代。那些不太擅长模式识别，因而不太容易生存和繁衍的个体将基因传下来的可能性要小一些。通过无数代自然选择，我们形成了寻找模式并对其做出解释的内在倾向——乌云会带来降雨，捕食者会发出声音，繁殖力强的成年人拥有漂亮的毛发。

遗憾的是，这种适用于远古祖先的模式识别技能并不能很好地适应我们的现代生活，因为我们面对的数据非常复杂，不是很容易解释。我们解释眼前事物的内在愿望导致了两种认知错误。首先，我们很容易被模式以及解释模式的理论所引诱。其次，我们紧盯着支持这种理论的数据，忽视与之相矛盾的证据。我们相信这些故事，因为它们与我们观察到的模式相符。一旦我们接受了这些故事，我们就很难放弃它们了。

当你在双骰赌桌上不断摇出七点时，你认为自己将继续保持连胜势头，因为你希望如此。当你不断投出两点时，你认为自己转运的时候到了，因为你希望如此。我们并没有认真地考虑过，骰子既不会记忆过去，也不会关心未来。它们是没有生命的；它们所具有的意义完全是我们这些满怀希望的人类赋予它们的。如果连胜势头持续，或者连败势头终止，我们就会更加相信自己想象出来的理论是正确的。如果这种情况没有出现，我们就会制造一些借口，以便坚持自己荒谬的信仰。

当运动员穿上没有洗过的幸运袜子，当投资者购买热门股票，当人们投入大笔资金购买不良资产、相信它们一定会触底反弹时，他们的行为和上面如出一辙。我们渴望使不确定的世界变得更加确定，渴望控制我们无法控制的事物，渴望预测那些无法预测的现象。如果我们穿着这种袜子取得了良好的表现，那么这一定是袜子的功劳。如果其他人购买这只股票赚了钱，那么我们购买这只股票也可以赚钱。如果我们运气不好，那么我们一定会转运，不是吗？秩序比混乱更加令人舒适。

由于这些认知错误，我们很容易受到各种统计性骗局的蒙蔽。当人们用毫无意义的模式来证明政府政策的后果、营销计划的不凡影响、投资策略的成功或者保健品的效果时，我们很容易认为这些模式是有意义的。由于我们在内心深处希望理解这个世界，因此我们形成了这种声名狼藉的、很难摆脱的弱点。

无论文，不生存

即使是受教育程度很高、应当具有冷静头脑的科学家也很容易受到模式的诱惑。在残酷的学术研究领域，聪明好胜的科学家一直在追求名誉和资助，以维持他们的事业。这种必要的支持是由他们在同行评议期刊中发表的成果决定的。“无论文，不生存”是大学生活中的一个残酷现实。

有时，在巨大的压力面前，研究人员甚至会撒谎和作弊，以实现个人的职业发展。为了生存，他们需要得到能够发表的结果，但他们的实验结果并不符合预期，这使他们感到沮丧；此外，他们还会担心其他人抢先发表类似的结果。因此，这些研究人员有时会对实验数据做手脚。毕竟，如果你相信你的理论是正确的，那么编造出证明这种理论的数据又有什么关系呢?

英国医生安德鲁·韦克菲尔德（Andrew Wakefield）制造的疫苗恐慌就是这种欺骗的一个严重案例。1998 年，韦克菲尔德和其他人在久负盛名的英国医学期刊《柳叶刀》上发表了一篇论文，称 12 名正常儿童在接种麻疹、腮腺炎和风疹（MMR）的疫苗以后患上了自闭症。实际上，在发表这篇论文之前，韦克菲尔德已经在一场新闻发布会上公布了他的研究成果，并且呼吁停止接种麻腮风三联疫苗。

许多家长看到了这方面的新闻报道，对于之前的例行程序产生了疑虑。他们的孩子患上自闭症的可能性似乎比感染麻腮风的可能性更加令人担忧；毕竟，后者在英国几乎已经绝迹了。一百多万名家长拒绝让他们的孩子接种麻腮风疫苗。

虽然我住在美国，但我和我的妻子也读到了这些新闻故事，并且产生了忧虑情绪。我们在 1998 年、2000 年和 2003 年生下了 3 个儿子，并在 2006 年生下了 1 个女儿，因此我们需要决定是否为他们接种疫苗。我们进行了调研，同一些医生进行了交谈。所有医生都对韦克菲尔德的研究持怀疑态度。他们指出，没有任何证据表明自闭症正在

变得更加流行；我们之所以听到更多的自闭症案例，是因为自闭症的定义近年来得到了拓宽，而且医生和家长对于自闭症症状的意识有所提高。另一方面，麻疹、腮腺炎和风疹是极具传染性的疾病，它们之所以在许多国家得到了有效的清除，恰恰是因为这些国家推行了常规免疫接种计划。如果我们的孩子不接种疫苗，不仅他们会面临危险，其他孩子也会陷入危险之中。此外，这项研究样本很小（只有 12 个孩子），而且作者似乎急于将其公之于众，这些都是很大的疑点。最终，我们决定为我们的孩子接种麻腮风疫苗。

不是只有我们遇到的医生产生了怀疑。一些人试图复制韦克菲尔德的结果，但他们并没有发现自闭症和麻腮风疫苗之间的任何关系。更糟糕的是，伦敦《星期日泰晤士报》记者布赖恩·迪尔（Brian Deer）在 2004 年进行了一项调查，在韦克菲尔德的研究中发现了一些可疑的反常之处。韦克菲尔德的研究似乎得到了一些律师的资助，这些律师希望接到针对医生和制药公司的利润丰厚的人身伤害诉讼。更加令人吃惊的是，韦克菲尔德本人显然正在计划推出一种替代性疫苗，他认为这种疫苗是安全的。韦克菲尔德的结论是否受到了这些利益冲突的影响?

韦克菲尔德声称自己并没有做出不道德的行为，但迪尔并没有停止挖掘。他发现了一些更加可恶的事情：韦克菲尔德论文中的数据与英国国民健康服务系统的官方医疗记录不符。在韦克菲尔德声称患上倒退型自闭症的 9 名儿童中，只有一个人被明确诊断出了这种疾病，还有 3 个人根本没有患病。韦克菲尔德声称 12 个孩子在接种麻腮风疫苗之前“是正常的”，但是其中 5 个人拥有发育问题的历史记录。

韦克菲尔德这篇论文的大多数共同作者很快撇清了自己与文章的关系。《柳叶刀》2010 年撤销了这篇论文，并表示，“论文中的说法显然是完全错误的，这是一件毫无疑问的事情。”《英国医学期刊》称韦克菲尔德的研究是“精心策划的骗局”，英国医学总会禁止韦克菲尔德在英国行医。遗憾的是，这个错误还是造成了一些损失。到目前为止，已经有数百名没有接种的儿童死于麻疹、腮腺炎和风疹，还有数

千名儿童处于危险之中。2011 年，迪尔获得了英国新闻奖，因为他对韦克菲尔德的调查“很好地纠正了一个错误”。我们只能祈祷韦克菲尔德的真面目能够像他的虚假警报一样得到媒体的大量报道，希望家长能够再次允许他们的孩子接种疫苗。

疫苗是一种注射到人体中的病原体，因此人们对疫苗的担忧存在一定的合理性，尤其是与孩子的安全有关的疫苗。不过，那些不合理的结论呢？人为制造的数据能够使我们相信那些明显具有荒谬性的观点吗?

迪德里克·斯塔佩尔（Diederik Stapel）是一位极为多产、成功的荷兰社会心理学家，以设计全面细致的调查方案著称，这些方案通常以研究生或同事为调查对象。奇怪的是，作为一名高级研究员，他竟然亲自开展调查，而且很可能是在只有他能够进入的学校里进行的。另一件奇怪的事情是，斯塔佩尔常常知道同事的研究兴趣，并且声称他已经收集到了同事需要的数据；斯塔佩尔可以提供这些数据，条件是他需要被列为研究报告的共同作者。

斯塔佩尔是几百篇论文的作者或共同作者，他还在 2009 年获得了实验社会心理学会的职业轨迹奖。2010 年，他当上了蒂尔堡社会和行为科学学院的院长。斯塔佩尔的许多论文具有争议性，但是至少看上去比较合理。不过，另一些论文已经超出了看似合理的界限。在一篇论文中，斯塔佩尔声称，凌乱的房间会加剧人们的种族主义倾向。在另一篇论文中，斯塔佩尔认为吃肉——甚至仅仅是考虑吃肉——会使人变得更加自私。（是的，我没有开玩笑！）

斯塔佩尔的一些研究生对于这些稀奇古怪的理论产生了怀疑，他们想要查看实际数据，以了解这些数据是如何支持他的理论的。令人失望的是，斯塔佩尔拒绝向他们展示调查数据。他们把这件事报告给了心理系主任。很快，斯塔佩尔承认，他的许多调查结果要么进行了篡改，要么完全是编造的。他解释说：“我想以很快的速度做出许多成果。”

2011 年，斯塔佩尔被停职，随后被蒂尔堡大学解雇。2013 年，斯塔佩尔放弃了自己的博士学位，并且撤销了 50 多篇含有虚假数据的论文。他还同意从事 120 小时的社区服务，并且放弃与 18 个月薪水相当的利益。作为交换，荷兰检察官同意不对他滥用公共研究基金的行为提起刑事指控，因为这些政府拨款主要用于支付研究生的工资，而这些研究生并没有犯下任何错误。与此同时，我们在吃肉和弄乱房间时也可以减少几分愧疚感。

伪造数据的另一个例子涉及超感知觉（ESP）测试。早期超感知觉实验使用心理学家卡尔·齐纳公爵（Duke Karl Zener）设计的一副纸牌。这副纸牌共有 25 张牌，包括 5 种符号：圆圈、十字、波浪线、方块和星星。在洗过牌以后，“发送者”依次查看每张纸牌，“接收者”需要猜测纸牌上的符号。

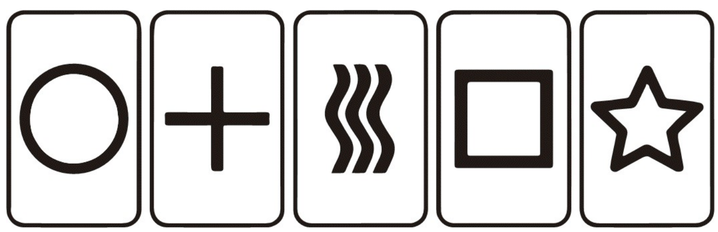


图 1.2　五种齐纳纸牌

一些怀疑者指出，接收者可以偷看纸牌，或者从发送者的行为中寻找细微的线索，比如快速扫视、微笑或者挑眉，从而获得较高的分数。沃尔特·J.利维（Walter J. Levy）是超感知觉先驱 J. B.莱因（J. B. Rhine）创立的通灵学研究所主任。为了平息上述批评，利维用计算机和非人类对象做了一些实验。在一项实验中，他把包含小鸡胚胎的鸡蛋放在孵化器里，并用一个电灯为孵化器加热，电灯的开关由计算机随机事件生成器控制。随机事件生成器开灯的可能性是 50%。不过，利维表示，小鸡胚胎能够对计算机产生影响，使电灯打开的时间超过一半。

利维的一些同事对这些“具有心灵感应能力的小鸡”产生了怀疑（我倒希望这是真的），而且对于利维在实验期间对设备的过度关注感到困惑。他们对计算机进行了改动，使之能够将结果秘密记录下来。然后，他们躲在一个隐秘的地方，对实验进行了观察。他们的担忧得到了证实。秘密记录表明，电灯打开的时间只占总时间的一半。而且，他们看到利维对设备进行了干预，将电灯打开的时间提高到了50% 以上。面对同事的质问，利维承认了自己的行为，并且辞去了职务。后来，他解释说，为了发表论文，他面临着很大的压力。

统计显著性膜拜

不过，我们最感兴趣的例子与虚假数据无关。这些例子涉及的做法更加微妙，更加常见。许多人对统计显著性非常关心。这是一种奇怪的宗教，它受到了研究人员近乎盲目地崇拜。假设我们想要通过检验了解，每天服用阿司匹林能否降低心脏病发作的风险。理想情况下，我们会对健康个体的两个随机样本进行比较。一个样本每天服用阿司匹林，另一个样本每天服用安慰剂——与阿司匹林具有类似外观、感觉和味道的惰性物质。这种检验应当是双盲检验：受试者和医生并不知道谁在哪个小组里。否则，病人报告“正确结果”（以及医生听到“正确结果”）的可能性也许会高于真实水平。

当研究结束时，统计学家就会介入进来。这里的统计问题是，在完全随机的情况下，两个小组的差异达到实际观测值的概率有多大。在大多数研究人员看来，小于 0.05 的概率具有“统计显著性”。如果数据中的模式仅凭运气出现的可能性不到 1/20，这种模式就会被视作具有统计说服力。章鱼保罗的数据具有统计显著性，因为它仅凭运气得到这种结果的可能性不到 1%。

在一项涉及 2.2 万名男性医生的阿司匹林研究的前五年，安慰剂组出现了 18 个致死性心脏病发作案例，阿司匹林组则只出现了 5 个同样的案例。在完全随机的情况下，这种巨大的差异发生的概率不到 1%。至于非致死性心脏病发作，安慰剂组有 171 个案例，阿司匹林组有 99 个案例。这种巨大的差异仅凭运气出现的可能性约为十万分之一。这些结果具有统计显著性，因此美国心脏学会目前建议心脏病发病风险较高的患者每天服用阿司匹林。

另一方面，没有发现统计显著性的结果有时比发现统计显著性的结果更加重要。1887 年，阿尔伯特·迈克耳孙（Albert Michelson）和爱德华·莫利（Edward Morley）测量了与地球运动方向平行和垂直的光速，希望找到二者之间的差异，以证明当时一种流行的理论。不过，他们并没有发现任何具有统计显著性的差异。他们的研究为爱因斯坦狭义相对论的提出和接纳奠定了基础。他们这项“失败”的研究为物理学革命做出了贡献。

再说一个更能说明问题的例子。我们将在后面的章节中讨论关节镜手术，这是一种针对膝骨关节炎的常规程序，每年会进行几十万次。最近的研究发现，这种手术带来的利益没有统计显著性，这一结论每年可以节省数百万美元资金，而且可以免去手术带来的不便和并发症风险。同许多发现罕见疾病疗法统计显著性的研究相比，这项没有在常见手术中发现统计显著性的研究显然具有更大的价值。

不过，一项针对心理学期刊的研究发现，在所有得到发表的检验结果中，97% 的结果具有统计显著性。当然，在研究人员进行的所有检验中，具有统计显著性的检验比例不会达到 97%。之所以出现上述结果，是因为编辑通常认为不具有统计显著性的检验没有发表的价值。

这个问题不限于学术领域。当商业或政府研究人员试图证明某个策略、计划或政策的价值时，他们往往觉得自己必须展示出具有统计显著性的经验性证据。不管是在哪里，研究人员都在追逐统计显著性，而这并不是一个难以寻觅的目标。在高速计算机和大量数据的帮

助下，寻找统计显著性是一件很容易的事情。如果你观察得足够仔细，你甚至可以在随机数据表格中发现统计显著性。

寻找统计显著性的一种方法是对多种理论进行检验，然后只宣布具有统计显著性的结果。即使只考虑毫无价值的理论，也会有 1/20 的检验表现出统计显著性。在海量数据和高速计算机的帮助下，在生成“可发表结果”的巨大压力下，无数毫无价值的理论得到了检验。成千上万的研究人员对无数理论进行检验，将那些具有统计显著性的结果记录下来，并将其他结果扔到一边。对于社会公众来说，我们只能看到这些统计工作的冰山一角。我们只能看到具有统计显著性的结果，看不到不具有统计显著性的结果。如果我们知道这些得到公布的检验背后隐藏着数百项没有得到公布的检验，而且知道对毫无价值的理论进行的检验中平均有 1/20 的检验能够得到统计显著性，我们一定会抱着更加怀疑的态度看待这些得到公布的结果。

比如说，制药公司会对数千种实验性药物进行检验。即使在设计良好的无偏研究中，我们也会发现数百种毫无价值的药物表现出具有统计显著性的效果——而这又可以带来巨大的利润。制药商很喜欢对更多的新药进行检验。不过，他们并不喜欢对得到批准的疗法进行重新检验，以查看最初的结果是否仅仅是一种巧合，是否属于 1/20 碰巧具有统计显著性的毫无价值的疗法。

当得到批准的疗法接受重新检验时，结果常常令人失望，这是一件意料之中的事情。约翰·约安尼季斯（John Ioannidis）在希腊约阿尼纳大学、马萨诸塞州塔夫斯大学医学院以及加利福尼亚州斯坦福大学医学院任职。（想象一下，他需要飞行多少公里，放弃多少睡眠时间！）在整个职业生涯中，约安尼季斯一直在警告医生和公众不要轻易接受没有以令人信服的方式得到重复的医学检验结果。在一项研究中，他考察了从 1990 年到 2003 年的 45 项备受尊重的医学研究，这些研究自称证明了对于各种疾病的有效疗法。其中，只有 34 项实验被人用规模更大的样本进行了检验，以复制初始检验结果。在这 34 项研究中，只有 20 项研究证实了初始结果（59%）。在 7 种疗法中，重新检

验得到的效果比最初的预测小得多；另外 7 种疗法的重新检验没有产生任何效果。总体而言，在 45 项研究中，只有 20 项研究得到了复制，而且这还是最受人尊重的研究！就在约安尼季斯发表这些令人不安的结果的同一年，他还写了另一篇文章，文章的标题很不讨好，叫做“为什么大多数得到发表的研究结果都是错误的”。

另一种获得统计显著性的方法是利用数据发现理论。统计检验的假设是，研究人员首先提出一种理论，然后收集数据，以检验这种理论，然后汇报结果——这种结果可能具有统计显著性，也可能不具有显著性。许多人颠倒了这种程序，他们仔细研究数据，以发现某种模式，然后编造出符合这种模式的理论。在数据中搜寻模式的过程令人愉快，而且激动人心，就像玩数独或者解决神秘谋杀案一样。这些人从各个角度考察数据，将数据分解成基于性别、年龄和种族的类别，丢掉妨碍模式的数据，寻找任何有趣的现象。当他们发现某种模式时，他们开始思考其中的原因。

当研究人员钻研数据、寻找模式时，他们会进行数百次显性或隐性检验。你可以站在他们的立场上进行考虑。首先，你将数据作为一个整体进行考察。然后，你分别查看男性和女性的数据。接着，你将儿童和成年人的数据区分开；然后将儿童、青少年和成年人的数据区分开；然后将儿童、青少年、成年人和老年人的数据区分开。接着，你尝试不同的年龄界限。你将老年人的范围设置为 65 岁以上。当这种做法失败时，你将这个数字调整为 55 岁、60 岁、70 岁或者 75 岁。最终，你总会发现某种模式。即使研究人员不对数据的每一种排列进行正式的统计检验，他们也可以进行非正式检验，即寻找看上去具有统计显著性的数据排列。如果我们知道研究人员在公布结果之前以一百种不同的方式对数据进行了考察，我们一定会抱着怀疑的态度看待这些结果。

这些做法——选择性报告和数据搜刮——被称为数据挖掘。通过数据挖掘发现的统计显著性只能体现出研究人员的耐心。在独立检验证实或拒绝结论之前，我们无法判断某种数据挖掘马拉松到底证明了

某种实用理论的有效性还是研究人员坚定的毅力。不过，通常情况下，这类检验并不会被人验证。毕竟，你无法通过证实他人的研究而成为明星；所以，为什么不把时间用于发现新理论呢？因此，通过数据挖掘得出的理论看上去很安全，既不会受到检验，也不会受到质疑。

许多重要的科学理论的确是人们为了解释他们所发现的模式而提出来的。例如，在 19 世纪，大多数生物学家认为，亲代性状的平均值决定了后代性状。例如，孩子的身高是由父母身高的平均值决定的，同时也可能受到环境的影响。

奥地利修士格雷戈尔·孟德尔（Gregor Mendel）在 8 年时间里对几万株豌豆进行了细致的研究。他对一些不同性状进行了考察，认为当时的混合理论是说不通的。当他对绿色种子的植株和黄色种子的植株进行异花授粉时，后代的种子要么是绿色的，要么是黄色的，没有黄绿色的种子。当他对圆粒种子植株和皱粒种子植株进行异花授粉时，后代的种子要么是圆粒的，要么是皱粒的，没有处于两者之间的种子。为了解释这些实验结果，孟德尔提出了“孟德尔遗传定律”，这个优雅的概率模型解释了性状的代际传递以及偶尔发生的隔代传递。他构造出了一种与数据相符的理论，为现代遗传学奠定了基础。

不过，数据挖掘还导致了数千种胡诌理论。我们如何区分正确理论与胡诌之间的区别？有两种有效的良方：常识和新数据。如果某种理论听起来很可笑，那么在看到压倒性的证据之前，我们绝不应该轻信这种理论；即使看到了压倒性的证据，我们也应当保持怀疑的态度。不同寻常的说法需要不同寻常的证据。遗憾的是，在这个年代，常识是一种稀缺品，许多诚实的研究人员用严肃的语气提出了一些愚蠢的理论。你知道吗？有人认为入选名人堂会使棒球选手的寿命预期减少五年。还有人认为一些中国人由于出生在“火年”而死于心脏病。本书稍后会对这些例子进行介绍。

第二种良方是新数据。当你通过搜刮数据编造出了某种理论时，用同样的数据对这种理论进行检验是一种不明智的做法。既然你编造

这种理论是为了迎合数据，那么这些数据当然会支持这种理论！应当使用没有被数据挖掘所污染的新数据对理论进行检验。

当你用新数据检验通过数据挖掘形成的理论时，得到的结果常常令人失望，这是意料之中的事情。用启发某种理论的数据对这种理论进行检验的做法显然具有误导性。如果某种理论对新数据的匹配程度不像原始数据那么好，这显然并不奇怪。

举个例子。我刚刚用左手小指弹开了桌子上的一枚 25 美分硬币，硬币落地时背面朝上。看到这一结果，我得出了一种理论：如果我用左手小指将桌子上的 25 美分硬币弹开，那么它在落地时总会背面朝上。毕竟，我的数据可以支持这种理论。这种理论显然愚蠢而毫无用处，但我们在后面几章将要详细考察的一些理论也是如此，这些理论虽然不那么容易被看穿，但它们与我的弹硬币理论具有相同的推导方式。如果死于癌症的孩子生活在输电线附近，那么输电线的电磁场（EMF）一定是导致癌症的元凶，不是吗？如果某种理论看上去比较合理，而你又不知道这种理论是通过观察数据得到的——比如观察地板上的硬币——那么你很容易不假思索地认为与数据相符的理论一定是正确的。毕竟，这种理论能够得到数据的证明！这是你应当抗拒的几种冲动之一。

幸运的是，我们能够抗拒这种冲动。我们能够克服我们的远古祖先在生存和繁衍的斗争中形成的偏好。我们可以摆脱数据的欺骗。

> 我们通常倾向于寻找模式并且相信我们所观察到的模式是有意义的。如果棒球选手在穿上新袜子以后在比赛中表现出色，那么他不应该换袜子。如果股票市场在 NFC 球队赢得超级碗以后表现出色，那么你应该首先观看比赛，然后再去投资。如果篮球选手连续四次投篮命中，这说明他的状态很好，很有可能在下次投篮时命中。如果心脏病患者在 1600 公

里以外的人发送治疗意念以后恢复正常，这说明远距离治疗是有效的。如果一项顾客满意度调查发现家里拥有三间浴室的人比家里拥有两间浴室的人更加热情，这说明前者才是企业的目标市场。如果一个国家在联邦债务很高时发生了衰退，这说明政府债务导致了衰退。在这本书中，我们将对几十个类似的例子进行反驳。

不要天真地认为模式就是证据。我们需要一个符合逻辑、具有说服力的解释，并且需要用新数据对这种解释进行检验。

第 2 章
不再神奇的超级畅销书

1791 年 12 月 26 日，查尔斯·巴贝奇（Charles Babbage）出生于伦敦。那是一个科技和社会流动性发生巨大变革的时代。巴贝奇对数学非常感兴趣，但他对数学和天文学表格中的人为计算错误感到非常沮丧。这些错误不仅仅是智力上的失败，它们还会带来严重的后果，包括致使船长将轮船驶入暗礁区或者其他危险地带。

当时，在高贵的英国人看来，关注法国数学家是一种不爱国的行为。不过，巴贝奇还是这样做了。他发现法国政府利用一种自动人工系统制作了一些数学表格。在这个系统中，高级数学家确定填充表格所需要的公式，初级数学家将这些公式简化成加减运算。例如，在计算 4×8 时，我们可以将其简化成加法：8+8+8+8=32。最底层的加减运算是由专业人员完成的，这些专业人员被称为“计算员”。

巴贝奇意识到，从理论上说，人们可以设计出以 100% 准确率进行加减运算的机器，从而消除人为错误。巴贝奇还知道两个德国人（威廉·席卡德和戈特弗里德·威廉·莱布尼茨）以及伟大的法国数学家布莱士·帕斯卡（Blaise Pascal）设计的计算机。帕斯卡的父亲是法国收税员，因此他在十几岁的时候设计了一台机械计算机，叫做“阿里特马提克”（又叫“帕斯卡利娜”），以便为他的父亲提供帮助。“阿里特马提克”是一个盒子，上面带有一些表盘，它们与盒子里隐藏的轮子相连。每个表盘上标有 0 到 9 十个数字。当个位表盘从 9 移动到 0 时，

十位表盘就会向上移动一个刻度；当十位表盘从 9 移动到 0 时，百位表盘就会向上移动一个刻度；依此类推。“阿里特马提克”可以进行加减运算，但是表盘需要手工转动。

巴贝奇将两种思想（将复杂公式转化成简单计算以及简单计算的自动化）融合在一起，设计出了一台机械计算机，可以完美地执行所有计算任务。作为巴贝奇的第一项设计，这台由铜和铁制造、由蒸汽驱动的庞然大物高达 2.4 米，重达 15 吨，包含 2.5 万个不同部件，被称为“差分机”。这台“差分机”可以进行多达 20 个数位的运算，而且可以将结果打印成格式化表格。在对这项设计进行了 10 年的小修小补以后，巴贝奇开始规划一台更加强大的计算机，叫做“分析机”。这项设计包含 5 万多个部件，使用穿孔卡片输入指令和数据，可以存储多达一千个 50 位数字。“分析机”有一个高达 4.5 米、直径为 1.8 米的圆柱形“工厂”，可以执行一个 7.5 米长的“存储器”发送的指令。这个“存储器”相当于现代计算机的内存，“工厂”相当于现代 CPU。

巴贝奇的核心原则非常合理，与现代计算机的工作原理类似。不过，考虑到当时的技术水平，他所提出的机械设计极为庞大，而且他需要不断面对资金限制以及所需部件短缺的困扰。尽管如此，他的宏大视野和对细节的关注仍然使人感到震撼，因此他的大脑（发明了计算机的大脑）一直被保存到了今天，并被展示在英国皇家外科医师学院里。

1991 年，在巴贝奇诞辰 200 周年之际，伦敦科学博物馆根据巴贝奇最初的计划制作了几台计算机，包括第二代差分机。这台机器和他预想的一样精确，可以进行多达 31 个数位的计算。2011 年，有人开展了一个名为“第 28 号计划”的私人非营利项目，用于制造巴贝奇的分析机，希望用巴贝奇领先时代一百年的超前思维激发我们的灵感。项目预计将在 2021 年巴贝奇逝世 150 周年之前完工。[书籍分 享V 信 zmxsh998]

由于巴贝奇的思想比他的时代领先了一个世纪，因此许多人对他的想法感到困惑。他在自传中回忆道:

> 有两次，（国会成员）问我：“请问，巴贝奇先生，如果你把错误的数字输入到机器里，机器会输出正确答案吗？”……我不太理解引发这种问题的混乱思想。

即使是在计算机已经普及的今天，许多具有良好出发点的人仍然坚持着这种错误观念：由于计算机不会犯下运算错误，因此它们是绝对可靠的。哈佛校友杂志 2014 年的一篇文章声称：“只要能够对足够多的信息进行量化，现代统计方法一定能够得出比一个人或者一小部分人更加正确的结论。”我们可以认为这种说法由于过于拐弯抹角而失去了意义，也可以认为这种说法是完全错误的。

事实上，如果我们让计算机去做一些愚蠢的事情，它也会忠实地遵从我们的命令。有一句俗语叫做“输入垃圾，输出垃圾”。也就是说，不管计算机多么强大，输出的价值仅仅取决于输入的质量。这种说法的另一个版本叫做“输入垃圾，输出福音”，它意味着人们往往过度相信计算机生成的结果，不会对输入进行审慎的思考。如果计算机的计算以不良数据为基础，那么它不会输出福音，只会输出垃圾。

遗憾的是，许许多多的人盲目崇拜基于误导性数据的计算。下面是几个例子。

去最好的学校

《纽约时报》华盛顿分部主任戴维·莱昂哈特（David Leonhardt）关于经济主题的作品获得了包括普利策奖在内的多个奖项。2009 年，他为《纽约时报》撰写了一篇专栏文章，对《穿越终点线》一书作了介绍。此书是由两位前大学校长威廉·鲍文（William Bowen）和迈克尔·麦克弗森（Michael McPherson）以及一位博士研究生共同撰写的，其中后一位作者负责的很可能是分析 68 所大学 20 万学生数据的繁重

工作。这本书的核心观点是，美国在“说服学生上大学”方面表现不错，但在“培养学生从大学毕业”方面则表现得不太理想。半数大学生无法毕业。

他们发现的第一个“罪魁祸首”是匹配不足：一些学生本来可以去毕业率较高的大学，但他们却选择了毕业率较低的大学。鲍文教授告诉莱昂哈特：“许多来自贫困家庭的学生本来有更好的选择，但他们并没有就读适合自己的学校，这使我感到极为震惊。”总体而言，在学分积点高于 3.5、大学入学考试成绩高于 1200 分的低收入准大学生中，大约有一半的人并没有选择更好的学校。

例如，90% 的密歇根大学学生可以在 6 年内毕业。在东密歇根大学，这个比例只有 40%。许多成绩足以进入密歇根大学的学生选择了东密歇根大学。要想从经济上解决这种匹配不足的问题，我们可以提高东密歇根大学的学费，或者降低密歇根大学的学费，以激励学生选择毕业率更高的学校。

实际情况并没有这么简单。这些数据都是垃圾，结论也不是什么福音。如果让这些所谓的“匹配不足”的学生就读密歇根大学，他们的毕业率可能会变得更低。研究人员认为这些学生是随机分配到密歇根大学或东密歇根大学的，就像医生被随机分配阿司匹林或安慰剂一样。不过，大学决策并不是科学实验。

当数据涉及人们的选择时（比如当人们选择上大学、结婚或者要孩子时），就会出现“自选择偏差”。在这种情况下，对于做出不同选择的人进行比较的做法是靠不住的。例如，我们常常听人说，大学毕业生的工资高于高中毕业生，似乎人们观察到的工资差异可以衡量上大学的财务回报。不过，大学毕业生之所以工资比较高，部分原因在于他们比不上大学的人更聪明，更有抱负。实际上，做出不同选择的人本身可能就是不同的。

类似地，上述“匹配不足”的观点也存在自选择偏差。学生不仅会选择上大学，他们还会选择上哪所大学。许多学生之所以选择东密歇根大学，而不是密歇根大学，也许是因为他们觉得自己很难从密歇根

大学毕业。这可能是正确的选择。毕竟，他们比我们更了解自己的情况。虽然东密歇根大学的总体毕业率较低，但我们无法从数据中判断出是否每个能够上密歇根大学但选择了东密歇根大学的学生都能毕业。

在我们通过观察人们的行为而收集的“观测性数据”中，自选择偏差非常普遍。人们可以选择自己的行为，因此他们的选择也许反映了自身的特点。要想避免这种自选择偏差，可以进行对照实验，即将人们随机分配到不同的小组里，然后告诉他们应该怎样做。不过，研究人员很少有权力仅仅由于自己需要实验数据而强迫我们去做自己不想做的事情，这对我们所有人来说都是一件幸运的事情。

要想进行有效的“匹配不足”研究，我们可以将那些同时被密歇根大学和东密歇根大学录取的学生随机分配到某一所大学里。然后，我们可以对两个小组的毕业率进行比较。实际上，在 20 世纪 60 年代，有人曾在密歇根州伊普西兰提市做过类似的事情，尽管这听起来有些不可思议。当时，在一项实验性学前教育计划中，人们根据抛硬币的结果决定是否允许家庭社会经济地位不高的黑人孩子进入学前班。这项研究发现，上过学前班的学生高中毕业和找到工作的可能性较高，被逮捕的可能性较低。这项实验证明了学前教育的价值，尽管它对输掉硬币游戏的孩子来说似乎有些残酷。

巧合的是，东密歇根大学就在伊普西兰提市。更加巧合的是，2012 年，东密歇根大学错误地向 7700 名学生（学生总数的三分之一）发送了电子邮件，称他们遭到了学校的开除。校长为这个“不可原谅的错误”道了歉。如果这些学生真的被开除，学校的毕业率将会变得更低。

莱昂哈特《纽约时报》专栏的一位读者发表了一条评论，认为教育质量比毕业率更加重要。（否则，我们可以抛弃论文、考试和出勤等令人讨厌的要求，直接向所有学生发放文凭，将毕业率提升至 100%。）莱昂哈特回复道：“大学毕业生的平均工资比大学辍学生高出 54%，所以学位看上去显然具有经济意义。”这又是一种自选择偏

差！选择上大学并通过努力学习获得学位的学生显然与大学辍学生存在系统性差异。

投票人数越多越好?

美国只有大约一半的合格选民在总统选举中投票。为了提高这个百分比，一个有趣的建议是将不投票的人的名字张贴在当地报纸或者互联网上，以便使他们感到羞愧。2014 年，《纽约时报》提出了一个更加激进的解决方案:

> 惩罚和监禁是导致人们不再参与政治（比如投票或政治实践）的两个原因……一份针对美国城市最边缘群体的大规模调查发现，在曾被警察拦截和盘问的群体中，投票概率降低了 8%；在曾被逮捕的群体中，投票概率降低了 16%；在被定罪的群体中，投票概率降低了 18%；在曾经遭到拘留或监禁的群体中，投票概率降低了 22%。

这段文字显然暗示了这样一条信息：如果减少逮捕和定罪，投票率将得到极大的提高。

在解雇警察之前，别忘了，这些数据都是观测性数据。也许，遭到盘问、逮捕和定罪的人并不是从人群中随机选择的。也许，他们曾经犯下罪行。也许，不投票的人更有可能犯罪，减少逮捕和定罪并不会提高投票率。

一醉方休

饮酒是一个长期困扰许多大学的问题，而且常常是辍学的一个驱动因素。即使校园内部禁酒，一些学生也会令人遗憾地在周围街区由于醉酒而遭到逮捕。学生为自己被逮捕而苦恼，教授为学生不学习而苦恼，家长为大学没有监督和保护自己的儿女而苦恼。

“代替父母”原则意味着大学在法律上拥有保护学生、不让他们做出错误决定的权力和职责。这个原则的应用经历了许多波折。不过，许多大学完全有理由感到担忧。一些学生和家长曾经控告学校没能尽到家长的义务。逮捕和死亡事件对入学申请也会产生负面影响，而入学申请又是大学的生命线。

1984 年，弗吉尼亚理工学院心理学教授 E.斯科特·盖勒（E. Scott Geller）在美国心理学协会的年度会议上发表了一篇研究论文，介绍了他在弗吉尼亚理工学院附近的三家酒吧里观察到的现象。（这比待在科学实验室里要有趣得多！）他发现，平均而言，以桶为单位要啤酒的人喝下的啤酒是以杯或瓶为单位要啤酒的人的两倍以上。他的结论是：“如果我们禁止使用啤酒桶，饮酒问题将会得到重大改善。”这一结论在全国范围内得到了发表。

盖勒发表过超过 350 篇研究论文，并且凭借将行为科学运用到现实生活中的出色表现获得了大学校友杰出研究奖和校友服务奖。不过，上面这项研究不在他做过的那些优秀研究的范围之内。根据常识，我们知道，这项研究中存在自选择偏差，因为以桶为单位要啤酒的人一定是想一醉方休，而且他们通常能够实现这个目标。他们也许面临着某种将花过钱的啤酒喝完的心理压力，但即使换成酒杯或酒瓶，豪饮者也不会改变自己大量饮酒的意愿。

在漫长而多产的职业生涯中，盖勒对大学饮酒问题进行了许多研究，在进行上述酒吧研究的 27 年以后，在 2011 年美国心理学协会会议上，他承认了一个显而易见的事实：许多大学生“想要一醉方休……我们在多项研究中指出，他们的意图会影响他们的行为。如果他们想要喝醉，那么你很难阻止这件事的发生。”

放下遥控器

电视频道如此众多，精彩节目却寥寥无几。照本宣科的真人秀；由缺乏才能的选手参与的才艺竞赛；某博士告诉人们（可能是演员）他们很优秀；某鉴定专家告诉人们（可能是演员）他们很愚蠢；喜剧节目不断出现令人讨厌的背景笑声。也许，看电视时关掉声音会让人减少几分痛苦?

美国最受尊重的广播和电视新闻工作者爱德华·R.默罗（Edward R. Murrow）曾说过：“电视的主要用途是欺骗和孤立我们，转移我们的注意力，逗我们开心。”这是 1958 年的事情，那是所谓的“电视黄金时代”。时至今日，情况并没有任何好转。

科学家很早就知道，看电视会使人的脑电波从敏捷而符合逻辑的贝塔波转变成放松而发散的阿尔法波。加里·拉尔森（Gary Larson）制作了一集优秀的动画片，名为《电视机产生之前的日子》，展示了一家人懒洋洋地躺在地板和沙发上、盯着一面空白墙壁的情景。

不管你观看的是空白的墙壁还是闪闪发光的“灯箱”，这种无须动脑的注视（常常伴随着心不在焉的吃喝）都会为你带来危害。2011 年，一群研究人员报告说，在澳大利亚，每天用 6 个小时看电视的人比从不看电视的人平均早死 5 年。简单计算一下，你会发现，在 25 岁以后，每看一个小时电视，寿命预期就会减少 22 分钟。你不仅在茫然的凝视中浪费了一个小时，而且额外失去了 22 分钟的生命。考虑到人们一生中看电视的习惯，研究人员得出的结论是，如果以法律形式禁止看电视，人们的寿命预期可以提高两年左右。

和之前一样，问题在于，这些数据是具有自选择偏差的观测性数据。也许，同那些拥有更加有趣的事情去做，而且自身健康状况足以支持他们这样做的人相比，选择整天看电视的人更加安静，更加抑郁，或者健康状况不是很理想。要想进行有效的实验，我们应该选择

两个小组，禁止一组看电视，强迫另一组每天看 6 个小时电视。如果我被随机分到禁止看电视的小组，我会感到难过的。

请原谅我的直白

在灾难性的滑铁卢战役之前的早餐上，拿破仑宣布："威灵顿是差劲的将军，英国士兵是一群乌合之众；我们将在午饭之前解决战斗。"对许多人来说，这个故事再一次证明了法国人令人讨厌而又毫无根据的傲慢。不过，美国运通和法国旅游局的一项研究发现，大多数过去两年对法国进行过一次以上休闲旅游的美国人并不认为法国人不友好。他们究竟是如何得到这个结论的呢?

下面是一种方法。假设一名研究员受雇证明法国是一个休闲旅行的好去处。这名研究员可以进行一项简单的调查，调查内容如下：

1. 过去两年你对法国进行过几次商务旅行?
2. 过去两年你对法国进行过几次休闲（非商务）旅行?
3. 你是否认为法国人不友好?

假设大多数只对法国进行过一次商务旅行的人认为法国人不友好。这也是他们从不自愿前往法国的原因。研究员将这些人的回答扔到一边。

假设大多数从未去过法国的人认为法国人不友好。这也是他们从不前往法国的原因。研究员将这些人的回答扔到一边。

假设大多数只对法国进行过一次休闲旅行的人认为法国人不友好。这也是他们从不再次前往法国的原因。研究员将这些人的回答扔到一边。

假设大多数对法国进行过不止一次休闲旅行的人不认为法国人不友好。太好了！这正是研究员想要的回答。法国旅游局可以根据这些回答进行广告宣传：大多数过去两年对法国进行过不止一次休闲旅行的美国人不认为法国人不友好。

从字面上看，这种说法是正确的，但它具有欺骗性。这种说法存在自选择偏差，因为选择重返法国的人几乎一定在上次旅行时玩得很开心。不过，这种说法向人们提供了错误的暗示：大多数曾经去过法国的人希望再次前往法国。

这种偏差隐藏在所有顾客满意度调查中。一家航空公司曾经打出这样的广告：在经常从纽约前往芝加哥的商业旅行者中，84% 的人更喜欢这家航空公司，而不是另一家航空公司。这个广告令人困惑的地方在于，在从纽约飞往芝加哥的旅行者中，只有 8% 的人选择这家航空公司。如果 84% 的旅行者更喜欢这家公司，为什么选择这家公司的旅客只有 8% 呢?

答案是，84% 这个数字所依据的调查对象是该公司从纽约飞往芝加哥的一架航班上的乘客。选择这家航空公司的旅行者更喜欢这家公司，这没有什么好奇怪的。真正奇怪的是，还有 16% 的旅客更喜欢另一家航空公司。不过，我们很难想象航空公司会打出这样的广告：“在乘坐我公司飞机的旅客中，16% 的人感到后悔。”

只有弱者留下来

自选择偏差并不是观测性数据的唯一潜在问题。20 世纪 70 年代，一项集体诉讼认为密西西比州戈斯市“佐治亚-太平洋锯木厂”在初始工作分配和晋升方面歧视黑人。该厂 50% 的劳动力是黑人，但大多数黑人员工进入工厂时从事的是级别最低的工作类型（“公用事业”），而且从未获得晋升。

在审讯中，工厂经理承认，电工长是工厂里唯一需要事先具备一定技能的工作岗位。其他所有岗位需要的技能都可以在工作中学习。公司的管理层也证实，雇用和晋升决策具有主观性，没有任何书面程序或具体标准。关于高级工作岗位黑人相对较少的原因，工厂经理的核心观点是："同承担责任、付出额外的辛苦、延长工作时间相比，黑人更喜欢从事不需要太多技能的工作。"

美国上诉法院第五巡回法庭并没有被这种具有侮辱性的观点说服。他们引用了美国最高法院在"黑泽尔伍德校区诉美国案"中的观点："在可以证明总体统计性差异的时候，这种差异本身可以……构成歧视模式或实践的初步证据。"他们还补充说："'差异性对待'案件通常要求提供歧视动机的证据，而作为一种便捷的歧视机制，公司官员缺乏标准的主观决策的证据可以满足这种要求。"

法庭的结论当然是正确的，但其统计性证据的某个部分却存在一个被所有人忽视的微妙缺陷。法庭展示了 6 年前进入公司入门级公用事业岗位并且领取相同工资的 11 名员工 1976 年的工资。1976 年，5 名白人员工的平均工资是 3.88 美元，6 名黑人员工的平均工资则只有 2.99 美元。这似乎是工资歧视的明显证据。不过，事情并没有这么简单。

这些数据属于回溯性数据。也就是说，这项研究选择了一群员工，然后回顾他们的情况，而不是考察他们未来的情况。在前瞻性研究中，研究人员选择一个样本，然后监督它在一段时间里的变化。例如，你可以每年对 1000 个人进行医学检查，以查看饮食与心脏病之间是否存在关联。你也可以每年对 1000 家公司进行观测，以查看分红政策与股票表现之间是否存在关联。相反，在回溯性研究中，研究人员选择一个样本，然后考察它的历史，比如考察 1000 名老年女性的医疗记录或者 1000 家公司过去的表现。

回溯性研究往往存在幸存者偏差。也就是说，当我们选择现在的样本并回顾过去时，我们只能看到幸存者。老年人的历史医疗记录中

排除了没有活到老年的个体。公司财务历史记录中排除了已经破产的公司。

对 1970 年入职的员工 1976 年的工资进行比较的做法是一种回溯性研究，因为它只考察了 1976 年仍然留在公司的员工，排除了 1970 年入职后自愿或被迫离开公司的人。假设公司在 1970 年以完全相同的工资雇用了 10 个白人和 10 个黑人，而且每个群体里有一半的人工作效率很高，一半的人工作效率不高。进一步说，假设在这 6 年时间里，在所有种族群体之中，效率很高的员工的工资提高了 30%，效率不高的员工的工资提高了 10%。如果所有 20 名员工 1976 年仍然留在公司里，白人和黑人的平均工资将会持平。不过，如果效率不高的白人员工被解雇，留下来的白人的平均工资将高于黑人的平均工资。如果效率很高的黑人员工离开公司，以追求更好的工作，那么白人的平均工资也将高于黑人的平均工资。我们并不知道这些事情是否真的发生过。关键问题是，回溯性研究无法完整地描述 1970 年入职的所有员工的情况。

许多观测性研究存在幸存者偏差。例如，健康维护组织在一项调查中发现，超过 90% 的成员对该组织感到满意。这里存在两种幸存者偏差，它们都在推高调查的满意度：一些人由于不满意而退出了这项计划，还有一些人离开了人世。

红狮酒店曾经打出占据整整一个版面的广告："在曾经入住红狮的商务旅行者中，98% 的旅行者还会再次选择我们。"原来，该公司对酒店里的顾客进行了一项调查，其中 98% 的人表示"他们在旅行时通常会选择红狮酒店"。显然，只住过一次红狮酒店的人并没有被这项调查包括在内。

下面是一个更加微妙的例子。有人对纽约市兽医院接收的从高层公寓楼坠落的 115 只猫咪进行了调查，发现从 9 层以及上楼层坠落的猫咪的死亡率为 5%，从不足 9 层的楼层坠落的猫咪的死亡率为 10%。根据医生的推测，这是因为从较高楼层坠落的猫咪能够将身体伸展开，形成一种降落伞效应。还有其他解释吗?

这个例子存在幸存者偏差，因为坠落后死亡的猫咪不会被送到医院。而且，许多猫主会放弃那些从高层坠落后奄奄一息的猫咪，而从较低楼层坠落的猫咪的主人往往更加乐观，更愿意花钱带它们去医院。

被损坏的飞机

二战期间，英国皇家空军（RAF）计划在飞机上安装厚钢板，以帮助它们抵抗德国战斗机和陆基高射炮的攻击。这些防护板极为沉重，无法用来覆盖整个飞机，因此英国皇家空军开始对结束轰炸任务的飞机进行调查，以收集飞机上子弹孔和弹片孔的位置数据。图 2.1 是一张示意图。如图所示，大多数弹孔位于机翼和飞机尾部，驾驶舱、发动机和油箱的弹孔则非常稀少——这说明防护板应当安装在机翼和飞机尾部。你同意这种观点吗?

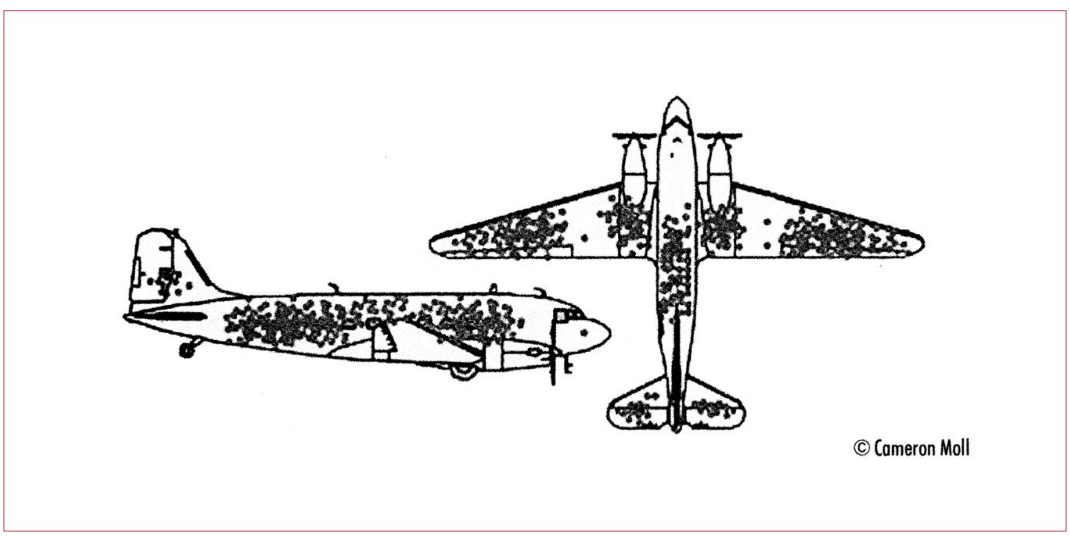


图 2.1　哪些部位需要加强防护?

逃到美国的匈牙利犹太人亚伯拉罕·瓦尔德（Abraham Wald）敏锐地意识到，这些数据存在幸存者偏差。返航飞机的驾驶舱和油箱之所以很少出现弹孔，是因为这些部位被击中的飞机无法存活下来并返回英国。返航飞机在机翼上出现弹孔的可能性更大，因为这些弹孔的伤害很小。瓦尔德的建议与最初的结论完全相反：他们应当加固没有弹孔的位置，而不是弹孔最多的位置。

这个办法发挥了作用。被击落的飞机数量大幅减少，安全返航并且可以重新投入战斗的飞机数量大幅增加。瓦尔德清晰的思路为战争的胜利做出了贡献。

畅销书的秘密

在撰写畅销书《从优秀到卓越》之前，吉姆·柯林斯（Jim Collins）及其研究团队花了5年时间，考察了1435家公司40年的历史，发现了股价表现优于平均水平的11家公司：

雅培	金佰利	必能宝
电路城	克罗格	沃尔格林
房利美	纽柯	富国银行
吉列	菲利普·莫里斯	

柯林斯对11家大公司进行了一番详细检查，发现了一些共同特点，并为每个特点起了一个吸引眼球的名字，比如第五级领导者——

具有谦逊的个人品质和职业驱动力、能够将公司打造成卓越企业的领导者。

柯林斯将他的工作描述成“寻找适用于一家组织的永恒而普遍的答案”。通过研究，他找到了他所寻找的宝藏：“只要认真使用我们所发现的思想框架，几乎任何组织都可以极大提高自身的境界和表现，甚至成为一家卓越的组织。”那些想要相信他的读者的确相信了他。《从优秀到卓越》的销量突破了 400 万部，而且进入了一些史上最佳管理类书籍的排行榜之中。

当然，问题在于，这是一项存在幸存者偏差的回溯性研究。下面是正确的研究方法。首先，选择这个 40 年时间段开始时存在的一组公司。它们可以是标普 500 指数中的所有公司，纽约股票交易所的所有上市公司，或者其他某个名单上的公司。重点是，这份名单应当始于 40 年前。接着，使用看上去合理的标准预测出表现优于其他公司的 11 家公司。这些标准必须具有客观性，不能参考这些公司未来 40 年的表现。先去查看哪些公司表现出色，然后再去预测哪些公司表现出色的做法是不公平的，也是没有意义的。这不是预测，只是回顾。

在选定 11 家公司以后，我们可以比较它们与其他公司在接下来 40 年时间里的表现。如果柯林斯是这样做的，那么 11 家公司中的某些公司无疑会令人失望。一些公司可能会破产。一些公司可能不会被他选中。这就是世界的残酷性。不过，这也将是一种公平的比较。

柯林斯并没有这样做。他在 40 年时间段结束时选择了 11 家成功的公司，以确保它们不会令人失望。柯林斯曾写道，他“直接根据数据进行了经验性推断，得出了所有概念”。他觉得他在宣扬自己的研究具有专业性，没有偏差；这些内容不是他编造出来的；他的一切观点都是通过数据推导出来的。

实际上，柯林斯承认了他对一些公司的表现优于其他公司的原因一无所知。他告诉人们，他对“根据数据推导理论”的危险性浑然不觉。为了支持这种理论在统计上的合法性，柯林斯找到了科罗拉多大学的两位教授。一位教授说：“你的框架中的概念随机出现的概率几乎

为零。”另一位教授说得更加具体，他说道：“这 11 家公司的员工表现出了你所发现的主要特点，而它们的直接比较对象并不具备这些特点。那么，随机找到这样 11 家公司的概率是多少呢？”根据他的计算，这个概率不到一千七百万分之一。柯林斯的结论是：“找到 11 个恰好具有我们所寻找的‘从优秀到卓越’模式的随机事件的可能性几乎为零。我们可以充满信心地得出结论：我们所找到的特点与从优秀到卓越的转变存在紧密的联系。”

我不知道这个一千七百万分之一的概率是如何计算出来的（我联系了这位教授，他说他不记得了），但我至少知道它是不正确的。这位教授的计算假定这五个特点是在查看数据之前确定的。事实并非如此，因此这种计算没有任何意义。正确的概率不是一千七百万分之一，而是 1。没错，是 100%。

假设我抓到了五张扑克牌：黑桃 3、黑桃 8、方片 8、红桃 Q、黑桃 A。这不是很神奇吗？抓到这把牌的可能性大约是三百万分之一，但它竟然出现在我的眼前！如果我在抓牌之前正确预测出了这五张牌，那才是一件神奇的事情。不过，如果我在抓到这把牌以后再去预测，那就一点也不神奇了。在我看到这把牌以后，拥有这五张牌的概率是 1，不是三百万分之一。

如果我们回顾历史上的任何一组公司，不管是最优秀的公司还是最糟糕的公司，我们总能发现一些共同特点。瞧，柯林斯选择的 11 家公司的名字里都有一个字母 i 或 r，有几家公司既有 i 又有 r。从优秀到卓越的关键是确保公司名称中有一个 i 或 r 吗？当然不是。

先选择公司、然后寻找共同特点的做法既不令人意外，也不令人感兴趣。真正有趣的问题是，这些共同特点是否能够预测哪些公司将在未来取得成功。

对于上面 11 家公司来说，答案是否定的。房利美的股价从 2001 年的每股 80 美元以上变成了 2008 年的每股不到 1 美元。电路城在 2009 年破产。在《从优秀到卓越》出版以后，其他 9 家公司在股市上

的表现非常平庸。从图书出版到 2012 年，五只股票的表现优于股市整体水平，六只股票的表现不如股市整体水平。

20 年前，另一部关于企业的畅销书做了一件非常类似的事情，这件事也存在完全相同的问题。这种循环现象给我们带来的真正教训是，写作这些书籍的作者和购买这些书籍的数百万读者没有意识到书中的本质缺陷。

世界顶级咨询公司之一麦肯锡公司要求两名默默无闻的咨询师汤姆·彼得斯（Tom Peters）和罗伯特·沃特曼（Robert Waterman）对几家成功的公司进行研究。彼得斯和沃特曼与麦肯锡的其他咨询师进行了交谈，列出了 43 家拥有良好声望和雄厚资金的公司。接着，他们与经理们进行交谈，并且阅读杂志上的故事，以寻找共同的主题。根据这项相当随意的研究，他们写出了一部极具影响力、极为成功的书籍《追求卓越》。此书列出了彼得斯和沃特曼在 43 家优秀公司中发现的 8 个共同因素，比如偏爱行动和接近顾客。这也是一项回溯性研究。我们无法知道“偏爱行动”（不管这种说法的含义是什么）的公司是否比其他公司更加成功，或者过去表现出色的公司能否在未来表现出色。

在这 43 家公司中，35 家公司拥有公开交易的股票。在这本书出版以后，15 家公司的表现优于股市的整体水平，20 家公司的表现不及股市的整体水平。柯林斯、彼得斯和沃特曼并没有提供任何证据证明他们所描述的特点是这些公司过去取得成功的原因。要想做到这一点，他们需要提供这些特点的理论证据，在不知道公司拥有或没有这些特点的情况下选择公司，然后根据某种事先确定的衡量标准监督它们的表现。这些作者并没有开展这样的工作。

对成功的企业、婚姻和人生进行回溯性研究的所有书籍都存在这个问题，包括成功企业、持久婚姻、活到百岁的方法/秘密/诀窍等。这类书籍存在固有的幸存者偏差。如果我们觉得自己知道成功的任何秘密，一个有效的检验方法是找到拥有这些特点的企业或个人，然后

查看他们未来 10 年、20 年或 50 年的表现。否则，我们仅仅是在考察过去，而不是在预测未来。

×如何轻松识破一本正经地胡说八道

我们会观察人们的工作、游戏和生活，而且我们会自然而然地根据我们看到的现象得出结论。我们的结论可能会失真，因为这些人所做的事情是他们自己选择的。我们观察到的特点可能并非源于活动，而是反映了选择这种活动的人的个性。

如果有人告诉我们，参与竞争性体育运动的孩子非常自信，我们不应该认为竞争性体育运动可以增强自信。也许，自信的孩子喜欢参与竞争性运动。如果有人告诉我们，华尔街的工作人员很有进取心，我们不应该认为华尔街可以培养进取心。也许，华尔街容易吸引具有进取心的人。如果获得普利策奖的新闻工作者和获得大学卓越奖项的人都会受到自选择偏差的欺骗，那么我们所有人都需要保持警惕。

我们会自然而然地根据我们看到的现象得出结论——工人的工资、受损的飞机、成功的公司。我们还应当考虑到我们没有看到的事情——离职的员工、没有返航的飞机、失败的公司。我们没有看到的数据可能和我们看到的数据一样重要，甚至更加重要。为避免幸存者偏差，应当从过去开始并向未来展望。查看 20 年前受雇的人员、出发参与轰炸任务的飞机、40 年前处于经营状态的公司——然后查看接下来发生了什么。

第 3 章
被误传的谋杀之都

我有一个朋友，名叫史蒂夫（Steve），他从加州理工学院退学，参加了陆军游骑兵。这是美国士兵中的一个精英群体，专门执行高危近战任务。游骑兵的信条中有这样一句话:“我的国家希望我比其他士兵走得更远、更快，战斗得更加英勇。”

史蒂夫的部队专门执行空降突击任务，他的膝盖在数百次跳伞经历中受到了很大伤害。一次，他的团队遭到伏击，被困在越南的几块稻田里。他们无法站起来作战或者奔跑，因此他们尽量压低身子，等待直升机过来将敌人驱离。幸运的是，一个绰号为“狂野比尔”的狙击手正驻扎在有利位置，以应对这种情况。比尔位于稻田的 300 米开外，但他经常可以命中三倍于这个距离的目标。当敌军士兵探头查看他们的藏身之处时，“狂野比尔”将他们一个一个地放倒。等到直升机赶到时，一部分敌军士兵已经死去，其余士兵全都逃走了。另一次，“狂野比尔”在一百米开外射杀了一名坐在汽车里的敌军将领。子弹威力极大，击穿了汽车的防弹挡风玻璃，穿过司机，击中了坐在后排的将军。

史蒂夫带着一对受伤的膝盖和一系列令人震惊的故事离开了游骑兵。在美国，他过着和平而活跃的生活。他参加体育运动，并且教人们进行水肺潜水。最终，他那对患有骨关节炎的膝盖使他放慢了脚步。骨关节炎是一种退行性疾病。面对无法进行大量运动的黯淡前

景，史蒂夫感到非常沮丧。一位医生告诉他：“你的年纪大了，不适合参加体育运动了。”这使史蒂夫感到很难过。

史蒂夫的情况并不是个例。几十年来，当我们站立、走路、奔跑、跳跃和舞蹈时，我们的膝关节一直在支撑我们超重的身体。因此，它们很容易撕裂，扭伤，出现一些浮游软骨。最常见的疗法是医院每年进行数十万次的关节镜手术。这种手术需要切出两个小型刀口，一个用于导入小型光导纤维摄像机，一个用于导入实施手术的小型仪器。外科医生需要移除碎片，然后对余下的结构进行修复、清洁、平整和修理。当这项耗资 5 000 美元的手术结束后，疼痛就会消失。至少，医生是这样说的。

史蒂夫接受了关节镜手术，并对结果大加赞赏。他反复告诉我，我应当接受这项手术，以修复我那娇嫩的膝盖。不过，我对此持怀疑态度。几十年来，人们并没有将关节镜手术与其他疗法进行比较。医生进行手术，患者表示他们感觉病情出现了好转。除此以外，我们还需要知道什么呢？首先，我们需要知道，他们的膝盖是否真的得到了改善？也许，患者之所以表示病情出现了好转，是因为他们认为自己应当感觉到病情的好转。永远不要低估暗示的力量。

为了避免接受一项没有必要的手术，我对这个问题进行了进一步的考察。要想判断手术是否有效，真正的科学方法是进行一项对照实验，对一部分随机选择的患者进行关节镜手术，并且不对其他患者进行这项手术。不过，有一个困难似乎无法化解：患者知道自己是否做过手术，这将影响他们对于自身感觉的反馈结果。

为了解决这个问题，在 20 世纪 90 年代，一项涉及 180 名退伍军人的对照实验设计了一个周密的策略。医生为对照组的患者切出两个浅浅的刀口，并且做出模仿关节镜手术的动作。这些患者并不知道他们是实验的一部分。在接下来的两年时间里独立评估老兵状况的医生同样不知道这项实验。研究的结论是，同接受假手术的对照组相比，真正接受关节镜手术的患者在任何时刻都没有感受到更少的疼痛，或者表现出更好的功能。显然，病人声称的疼痛缓解完全来自安慰剂效

应（人们相信医学的力量，希望看到治疗方法带来积极的效果，即使这种治疗方法没有任何医学价值）。

这项研究2002年发表在《新英格兰医学期刊》上。6年后，发表在同一份期刊上的另一项研究证明，对于膝骨关节炎患者来说，同只接受药物和物理治疗的对照组相比，接受关节镜膝盖手术、药物治疗和物理治疗的患者在疼痛、僵硬度和物理功能方面并没有表现得更好。现在，许多医生建议病人放弃这项手术。我的医生是一位运动医学专家，他向我提出了同样的建议。我相信他的观点。

胃冷冻是一种治疗胃溃疡的奇特疗法，目前已经丧失了信誉。这种疗法也经历了一个与治疗膝盖问题非常类似的过程。胃溃疡有时会使人感到极为痛苦。过去，人们常常通过物理切除胃酸来源的方法治疗这种疾病。一位具有发散性思维的医生想，既然他们可以用冰块减少踝关节扭伤和其他外伤的疼痛，那么他们也许可以用冰块使病人的胃部失去知觉，从而减少疼痛。不过，让病人吞下几十个冰块的方法不仅会使病人感到很难受，而且缺乏效率，因为你无法保证冰块与溃疡部位的持续接触。

人们提出了解决方案：将气球插入到溃疡患者的胃部，然后通过气球泵入过冷液体。这显然比手术更加便宜，更加安全，尽管它的效果可能不像手术那样持久。20世纪50年代的实验表明，这种古怪的想法实际上是有效果的，因为患者报告说，他们的胃酸分泌量出现了减少，胃部疼痛得到了缓解。这些结果发表在很有声望的《美国医学协会期刊》上。在此后的几年时间里，人们一直在用胃冷冻疗法缓解溃疡患者的疼痛。

和治疗膝盖问题的关节镜手术一样，胃冷冻疗法缺乏相应的对照组，因此我们无法知道这种疗法是否真的有效果。在被问及胃部疼痛时，患者可能倾向于给出他们所认为的正确答案。

和之前一样，我们需要一项对照实验，将过冷液体泵入一组随机选择的患者体内，将与体温相当的液体泵入另一组患者体内。当然，我们不能把我们所使用的液体种类告诉患者。

当人们最终结束这项实验时，得到的结果令人吃惊。在接受胃冷冻治疗的患者中，34% 的患者表示病情出现了好转；在接受与体温相当的液体的患者中，这个比例是 38%。又是安慰剂效应！随后的研究表明，胃冷冻没有任何真正的效果，医生从此不再将冰冷的气球塞进人们的喉咙。

正像这些例子表明的那样，设计良好的实证研究通常涉及比较。不过，正像下一个例子说明的那样，比较并不总是公平而有效的。有时，人们会对苹果和李子干进行比较。

马萨诸塞州谋杀之都

如果道琼斯工业平均指数一天下降 100 点，这个数字是多还是少呢？要想正确看待这种变化，我们可以计算下跌百分比。如果道指当天开盘时是 1000 点，那么 100 点的下跌对应于 10% 的比例（令人恐慌）。如果道指当天开盘时是 10000 点，那么 100 点的下跌对应于 1% 的比例（令人不快）。股票价格的百分比变化是一种有用的信息。不过，一些百分比变化也会具有误导性，比如当人们将某种小型事物的百分比变化与某种大型事物的百分比变化进行比较时。

韦尔弗里特是马萨诸塞州的一座小镇，以牡蛎、艺术家和宁静著称。不过，波士顿一份报纸报道说，韦尔弗里特当年的谋杀率是马萨诸塞州之最，每 10 万名居民中就有 40 起谋杀案——这个数字是波士顿的两倍多，后者每 10 万名居民中只有 17 起谋杀案。这篇报道使人感到非常吃惊。一名困惑的记者对这件谋杀统计谜案进行了研究，发现韦尔弗里特的所有警察都没有听说过这座小镇发生的谋杀案，包括一位在韦尔弗里特生活了 50 年的警察。

不过，一名被控在 30 公里外犯下谋杀案的男子在韦尔弗里特警察局自首，这个案件被错误地统计成了发生在韦尔弗里特的谋杀案。韦

尔弗里特只有 2491 名居民，因此一项被错误统计的案件变成了每十万名居民中的 40 起谋杀案。相比之下，波士顿发生了 98 起谋杀案，相当于每十万名居民中发生 17 起谋杀案。

这个谋杀之谜显示了当基数很小时一个统计意外所产生的巨大影响。在波士顿，一项错误记录的谋杀案对谋杀率的影响很小。在韦尔弗里特，一项错误记录的谋杀案可以将一个以牡蛎和艺术家著称的小村庄与底特律画上等号。对于基数较小的情形，一种处理方法是使用许多年的数据，以获得更大的基数。在过去 50 年时间里，韦尔弗里特只有一起被错误记录的谋杀案，或者说没有谋杀案——两种说法都可以证明，这的确是一个和平的小镇。所以，不要担心，把牡蛎递给我吧。

请在我家后院开一座采石场

蒂梅丘拉是一个南加州社区，与洛杉矶、圣迭戈和奥兰治县之间的距离大致相等。虽然蒂梅丘拉位于内陆山谷之中，但它距离太平洋只有 20 公里。每天，海风通过一个山间空隙带来凉爽的空气，使蒂梅丘拉山谷形成了温和的地中海气候。

由于拥有怡人的气候、乐观进取的精神以及支持商业的态度，蒂梅丘拉的人口从 1980 年的不到 2000 人增长到了 2010 年的 10 万人以上，家庭年收入中值超过 8 万美元。这个小镇还拥有 52 万平方米的体育设施，包括 10 个棒球场（5 个带有照明设施），两个带有照明设施的橄榄球/足球场，一个游泳池以及体育馆和社区中心。

晴朗的白天和凉爽的夜晚不仅可以吸引人口，而且适合生产葡萄酒。蒂梅丘拉山谷目前拥有 30 多家酿酒厂。这里还有数百家古董商店、特色商店和餐厅，9 个高尔夫球场，以及加州最大的赌场。这里有热气球比赛、汽车展览、爵士音乐节以及电影节。在不到 30 年的时

间里，这个很小的小镇已经升级成了一个绝佳旅游目的地，每月有 7 万人入住蒂梅丘拉酒店，1/6 的当地居民从事旅游业。

2005 年，美国最大的矿业和建筑公司申请在蒂梅丘拉附近建造采石场的许可证——实际上，这是一个超级采石场，面积相当于 17 个足球场，深度与帝国大厦的高度相同。这个采石场每天将动用 4500 公斤炸药，以便将岩石从山上分离出来，然后将其粉碎成石子和沙土（叫做“骨料”），用作混凝土、沥青以及其他建筑材料。除了粉碎设施，他们还将建造两个沥青厂、一个混凝土厂以及一个碎石回收厂。采矿和加工每天持续 20 小时，每周进行 6 天，每年将生产 500 万吨骨料。装货和运输每天持续 20 小时，每周进行 7 天，每天将有 1600 辆卡车进出该地。

根据计划，这个采石场位于海风将凉爽空气吹入蒂梅丘拉山谷时经过的山间空隙。因此，居民担心这个超级采石场将会影响城市的经济和房产价值。如果海风将灰尘、噪音、刺鼻的味道以及其他污染物带进这个美丽的城市，旅游业者和房屋所有者该怎么办呢?

不必担心，矿业公司说。该公司请当地一位优秀经济学家准备了一份经济分析报告。报告称，这个采石场每年将为居民带来 1.72 亿美元的好处，而且不会带来任何成本。矿业公司的真实身份是圣诞老人，他将为那些仍然相信圣诞老人的人带来上亿美元的礼物。

如果事实真如该公司所说，那么各个城市将会为采石场的选址展开激烈的争夺。他们将不会为了吸引专业体育团队而向他们提供税收优惠、低息贷款和基础设施。相反，他们将会把这些激励政策提供给矿业公司。“这是 1 亿美元。请在我们城市开矿吧。”

不过，他们并没有这样做。

就连该公司雇用的经济学家也承认这一点：“骨料开采作业几乎总会遇到反对声音。”为什么人们要反对没有任何附加条件的上亿美元利益呢？是他们太幼稚，还是受雇咨询师的估计存在严重误导性？你觉

得呢？让我们研究一下这位咨询师对收益和成本的估计，看看我们能够发现什么。

关于经济收益的计算出现在几张电子表格之中，这些表格充满了令人头疼的细节，比如碎石价格和去除水中细颗粒物所需要的絮凝剂用量。所有这些细节形成了一种看似严格的经济分析，同时掩盖了一个事实：在电子表格最后得到的1.72亿美元中，有1.50亿美元并不是当地收益，而是公司的年利润。没错，由于公司销售的是蒂梅丘拉的“土地”，因此这位经济学家将1.50亿美元的公司利润统计成了当地居民的收益。实际上，除了看到一家公司变富所带来的某种间接快感，这种利润不会为当地带来任何利益。

剩余2200万美元的预期利益来自新采石场的99个工作岗位，其中大部分是将骨料运往建筑工地的卡车司机。这里同样存在问题。计划中的蒂梅丘拉采石场位于里弗赛德县靠近圣迭戈边境的位置。公司预计，蒂梅丘拉采石场生产的骨料60%将被运往圣迭戈县，这将减少圣迭戈对更加遥远的里弗赛德县采石场的骨料需求。

该公司的咨询师表示，骨料总产量不会发生变化。一位咨询师是这样说的：“新采石场所做的事情仅仅是替代原有采石场的供应。新的采石场不能也不会导致更多骨料被开采、运输和使用。”

在这里，咨询师的观点是正确的。骨料总产量是由需求决定的——也就是正在建设的住宅、商店和道路的数量。新的采石场只能改变骨料的生产地点。如果蒂梅丘拉的产量是500万吨，这将意味着里弗赛德县其他地区的采石场减产500万吨。该公司自身的说法暗示了产量或工作岗位不会增加，只会从现有采石场转移到新的采石场。

因此，公司声称为当地居民带来的1.72亿美元利益完全是空中楼阁——它只是建筑公司的1.50亿美元利润以及99个从现有采石场转移到新采石场的工作岗位而已。“该公司不会带来的”不是成本，而是利益！

那么。经济学家所说的“没有任何成本”呢？蒂梅丘拉居民对房产价值的担心是杞人忧天吗？另一位咨询师声称，在 70 公里外的科罗纳市，矿山生产与房产价值之间出现了“直接的正相关”。城市领导人显然应当欢迎超级矿场选择他们这座美好的城市——因为城市房产价值将受到“超级影响”。

这种说法令人无法容忍。当任何两样事物随时间增长时，它们之间可能没有任何因果关系，但它们仍然具有统计相关性。表 3.1 显示了美国的啤酒销量和已婚人口的数量。二者的相关性达到了惊人的 0.99。面对这种强烈的相关性，我们能否认为饮酒会导致婚姻？或者倒过来，我们能否认为婚姻会导致饮酒？

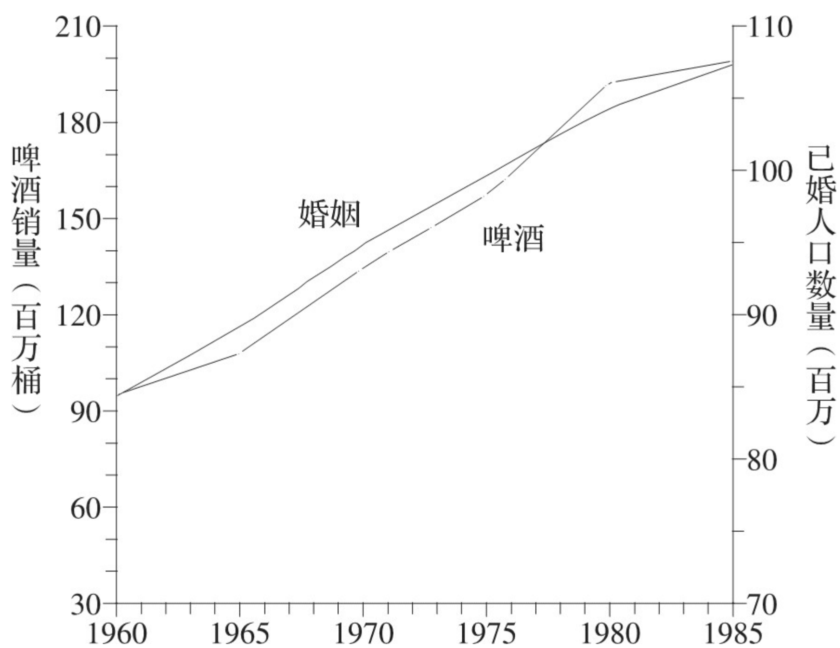


图 3.1 饮酒与婚姻

当然，一定会有一些已婚人士同意这种结论。不过，正确的解释是，当人口随时间增长时，啤酒消费量也会增长。类似地，婚姻、婴儿、汽车、鞋子、大学入学率、心脏病以及其他许多事物也会增长。

穿鞋会导致婴儿数量增长吗？开车会导致心脏病吗？这些事物的增长源自人口的增长，它们之间不一定存在任何关系。

矿山生产和房产价值也是同样的道理。二者都在随时间增长，它们之间不一定存在因果关系。公司出钱要求咨询师证明一些事情。当咨询师发现某种表面上的证据时，他们就会死死抓住这种证据不放。

我们都会做这样的事情。这种事情极为常见，甚至拥有一个名称：证实性偏差。我们认为自己能够预测橄榄球比赛、选举或者股票选择的结果。我们高估了自己的预测能力。如果我们的预测是正确的，这将证实我们的确是聪明人。如果我们预测失败，这仅仅是运气不好而已——裁判不公、投票率不高、其他投资者不理性。

当然，咨询师通常会受雇寻找某些事情的证据，这些事情可能连他们自己都不相信。真的有人认为采石场会提高附近住宅的价值吗？这种说法甚至无法通过表情测试。你可以将这种荒谬的说法讲给你的朋友听，看看你能否忍住笑声。

不过，只要价格合适，一些人愿意寻找能够证明荒谬观点的证据，然后装模作样地向人们证明这种观点。在寻找令人信服的证据时，他们常常会找到巧合的相关性，而那些随着时间和人口的增长而增长的数据往往具有巧合的相关性。

这正是矿业公司所做的事情。他们甚至用周围城镇具有误导性的统计数据支持自己的观点。另一位咨询师（他们雇用了许多咨询师）指出，加州科罗纳市的采石场已经存在了几十年，那里的房价与（目前）没有采石场的蒂梅丘拉具有几乎相同的增长率。因此，采石场不会抑制房产价值。

考虑这种观点背后的逻辑。如果牛排和土豆的价格具有相同的增长率，这是否意味着牛排的价格和土豆一样？假设两幢相同的住宅位于具有同等吸引力的位置，二者相距八公里。两幢住宅的市场价值相等，而且每年稳定地增长 3%。1980 年，一幢住宅旁边开了一家采石

场，其价值立即下跌了 20%。之后，两幢住宅的价值仍然每年增长 3%，二者的市场价值始终维持着 20% 的差距，如图 3.2 所示。

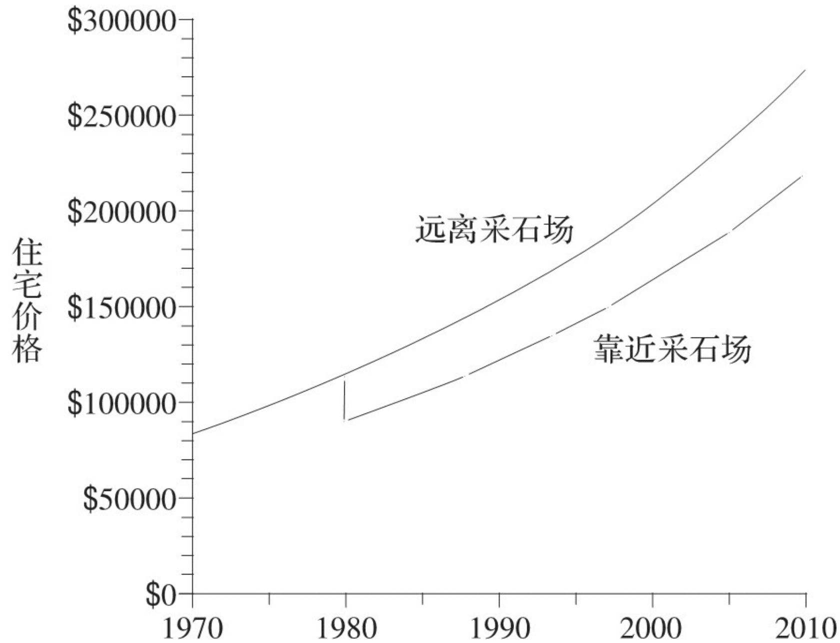


图 3.2　住在采石场旁边

正如矿业公司所说，靠近矿山的住宅与八公里外的住宅具有相同的价格增长率。这种平行的变化能否告诉我们采石场对附近住宅的价值具有积极影响、消极影响或者没有影响？答案是否定的。要想回答这个问题，唯一的方法是查看哪幢住宅拥有更高的价格。

每一个优秀的魔术师都知道，要想骗过观众，关键在于转移观众的注意力。这位咨询师希望通过谈论房价的变化使我们忘记真正的问题：靠近采石场的住宅是否拥有更低的价格？答案当然是肯定的。

我们很少能够找到两幢完全相同的房子，一幢靠近采石场，一幢远离采石场。不过，我们可以使用一个包含各种住宅特点的统计模型，包括建筑面积、浴室数量、与便利设施（比如好学校）的距离、与扰民设施（比如矿山）的距离。利用许多住宅的数据——有的住宅面积是 180 平方米，有的是 240 平方米；有的住宅有两间浴室，有的

有三间浴室；有的住宅靠近矿山，有的远离矿山——我们可以估计出增加 10 平方米面积的价值，增加一间浴室的价值，距离矿山远近的价值。

2006 年，一份针对俄亥俄数千幢住宅的研究做了同样的工作。研究结果显示距离采石场 1.6 公里的住宅价值下降了 15%，距离采石场 3.2 公里的住宅价值下降了 9%——如果你说服矿业公司将采石场开在你家后院，你的住宅价值并不会因此而提高。

在这个强烈支持商业的社区里，两家支持商业的机构——蒂梅丘拉山谷旅游局与商会——听取了矿业公司的报告，但是并没有被说服。这两个团体强烈反对建造采石场。

公司花了七年时间和数百万美元资金，试图使他们的采石场获得批准。不过，在阅读了几千页分析报告、听取了相关市民和双方专家几十个小时的证词以后，里弗赛德县计划委员会和行政管理委员会拒绝了这项采石场提案。他们显然并不相信圣诞老人。

×如何轻松识破一本正经地胡说八道

比较是实证研究的生命线。在与某种替代方案进行比较之前，我们无法确定某种药物、疗法、政策或策略的有效性。不过，请当心肤浅的比较，包括对于大数和小数百分比变化的比较，对于除了随时间增长以外没有其他共同点的事物的比较，以及对于无关数据的比较。这些比较就像苹果和李子干之间的比较一样。

第 4 章
新的经济学上帝

计算机可以迅速而正确地完成计算任务，但它不会考虑这些计算的意义和合理性。计算机只会做人们要求它去做的事情。

有时，错误的问题会导致错误的答案。如果我想知道 169 的平方根，实际却让计算机计算 196 的平方根。计算机会告诉我，答案是 14。对于我所提出的问题来说，这个答案是正确的。对于我想提出的问题来说，这个答案是错误的。这类错误被称为“计算误差”，但它们实际上是人为误差。遗憾的是，一些计算误差具有极为严重的后果。

1998 年，美国宇航局发射了一艘名为“火星气候探测者”的宇宙飞船。宇航局的计划是，当探测者号接近火星时，它将被发送到距离火星 160 多公里的轨道上。在那里，它将安全地记录天气数据。不过，这项耗资 3 亿美元的太空任务最终失败了，因为陆基软件以磅（“1 磅约为 0.45 千克——译者注”）每秒为单位测量推力，探测者号使用的则是公制单位牛顿，但是工程师忘记了单位转换。当探测者号接近火星时，发动机开始点火，试图将宇宙飞船送到围绕火星的轨道上。不过，由于这种计算错误，探测者号被推进到了距离火星 60 公里的地方，在大气中烧毁了。三亿美元打了水漂，而这仅仅是因为一个打印错误。

2012 年，发生了一个完全不同的、更为昂贵的错误。摩根大通英国分部在“伦敦鲸”崩溃事件中损失了 50 到 100 亿美元。“伦敦鲸”指的

是押下巨大赌注的那个交易员。当时，由于一个愚蠢的错误，银行的风险被严重低估。在衡量风险时，人们需要计算一个简单的平均值；例如，6 和 10 的平均值是(6+10)/2=8。程序员忘记了除以观测值数量的步骤；因此，平均值被错误地计算成了 6+10=16。结果，计算出来的风险值比实际风险小得多。摩根大通认为伦敦鲸的交易相对安全，但是这些交易事实上非常危险，并且最终酿成了灾难。

人们不仅会为计算机提供错误的指令，有时还会提供错误的数据。美国国会联合经济委员会曾报告说，美国最富有的 0.5% 家庭拥有国家 35% 的财富。同 20 年前的 25% 相比，这是一种令人震惊的增长。政客们发表了演讲（真是稀奇）。报纸用耸人听闻的标题报道了这个故事。联合经济委员会主席、威斯康星州民主党人戴维·奥贝（David Obey）表示：“这说明富人变得更加富有”。他还指出，之前的最高纪录是 1929 年的 32%。奥贝总结道：“这项研究彻底驳斥了‘美国需要为富人提供更多激励’的观点。”杰出的经济学家约翰·肯尼思·加尔布雷斯（John Kenneth Galbraith）发出了不祥的警告：20 世纪 20 年代财富的过度集中是导致大萧条的原因之一。普通人买不起商品，公司请不起工人生产人们买不起的商品。纽约社会研究新学院经济学教授戴维·M.戈登（David M. Gordon）警告说，“美国的民主制度已经到了最危险的时刻，这是过去 50 年不曾有过的事情——可能也是共和国成立以来不曾有过的事情。”

华盛顿的一些怀疑者对计算过程进行了检查，发现这种增长几乎完全来自一项错误记录：某个家庭的 200 万美元财富被写成了 2 亿美元。在调查比例接近 50% 的少数富人中，这一错误明显提高了平均财富水平。修正后的数据显示，最富有的 0.5% 家庭拥有美国财富的 26.9%，而不是 35%。

在这个错误被发现以后，奥贝发布了一份奇怪而虚伪的声明：“我们很高兴地得知，财富在超级富豪手中集中程度的增长不像最初表现得那么大。”其他经济学家对之前的声明闪烁其词，但是拒绝放弃他们的结论。

这个错误得到了及时纠正，没有对经济政策造成任何影响。不过，另一件差错的纠正则没有这么迅速。

政府债务临界点

2010 年，两位哈佛教授卡门·莱因哈特（Carmen Reinhart）和肯·罗戈夫（Ken Rogoff）发表了一篇论文，认为当联邦政府债务相对于国内生产总值（GDP）的比例超过 90% 时，国家的经济增长就会受到拖累。在这种情况下，国家需要花费 90% 的产值偿还政府的债务。

这听上去很合理，但是经不起推敲。政府并没有在短期内偿还债务的迫切理由。当我们购买住宅时，抵押贷款比我们的年收入大得多。那又怎样呢？真正重要的问题是我们的收入是否足以支付月供，而不是我们的收入是否足以立即付清贷款。政府也是一样的道理。而且，政府的负担要轻一些，因为它们可以在必要时开动印钞机。此外，90% 这个比例有什么特别之处吗？为什么不是 80% 或者 100% 呢？你无法找到一个可信的理由证明 90% 这个神奇的数字是经济增长和衰退之间的分界线。

不过，这些教授的确认为 90% 的债务与 GDP 之比是一个临界点，如果超过这个比率，国家将陷入衰退。而且，人们接受了这种观点，因为这毕竟是哈佛大学教授的结论。表 4.1 列出了两位教授最有说服力的证据。莱因哈特和罗戈夫研究了 20 个发达国家（澳大利亚、奥地利、比利时、加拿大、丹麦、芬兰、法国、德国、希腊、爱尔兰、意大利、日本、荷兰、新西兰、挪威、葡萄牙、西班牙、瑞典、英国和美国）。他们计算了每个国家每年债务与 GDP 的比率（以百分数表示）以及扣除通胀因素后的 GDP 增长率。根据债务与 GDP 的比率，这些数据被分成了四个类别。

表 4.1　1946～2009 年 20 个发达经济体的债务和 GDP 增长

	政府债务与 GDP 的比率			
	30% 以下	30% 到 60%	60% 到 90%	90% 及以上
平均 GDP 增长率	4.1	2.8	2.8	-0.1

表 4.1 显示，债务与 GDP 之比的增长会降低经济增长率。如果这个比率超过 90%，经济将陷入衰退之中，GDP 将会下降，经济总量将会萎缩。莱因哈特和罗戈夫的结论是，“至少，这意味着传统债务管理问题应当成为公共政策的首要关注点。”

作为支持紧缩的强有力证据，这项结论获得了全球的关注。许多国家的财政强硬派用罗戈夫和莱因哈特的研究来证明政府应当减少开支，提高税收，以平衡预算，甚至获得盈余。这样，政府债务才能得到偿还。

在美国，共和党 2013 年的财政预算指出：

> 经济学家肯·罗戈夫和卡门·莱因哈特完成的一项著名研究证实了这个常识性结论。这项研究发现了决定性的现实证据，证明超过 GDP90% 的总债务（指政府欠下的所有债务，包括政府信托基金持有的债务）对经济增长具有重要的负面影响。

谁能对常识和决定性现实证据提出异议呢？

在这种宣传的影响下，2013 年，《华盛顿邮报》的一篇社论发出了不祥的警告：“债务与 GDP 之比可能会持续攀升——并且危险地维持在被经济学家们视作威胁持续经济增长的 90% 关口。”在这里，两

位经济学家的一项研究似乎被《华盛顿邮报》描述成了所有经济学家的普遍意见。

事实上，许多经济学家从不认为存在一个 90% 临界点。大多数经济学家的共识是，当一个国家的经济仍然没有从 2007 年 12 月开始的大萧条中恢复过来时，为了减少政府债务而削减开支和提高税收的做法反而是错误的。许多人对罗戈夫和莱因哈特的研究感到极为怀疑，但是他们很难找到任何严重的错误。

一个问题是，美国债务与 GDP 的比例只在四年时间里超过 90%。美国经济在这四年里出现了萎缩，但那是非常特殊的年份：1946 ~ 1949 年，正好是二战结束以后。政府债务之所以异常偏高，是因为政府为了支持战争而大量借贷。战后的衰退是由政府军事开支的下降引起的，不是由战争期间积累的债务导致的。显然，我们不应该认为这四年证明了政府债务会导致衰退。相反，这四年实际上证明了财政强硬派倡导的大幅削减政府开支的做法会导致衰退。

不过，世界各地的财政强硬派都在使用莱因哈特和罗戈夫的研究支持英国保守党领导人戴维·卡梅伦（David Cameron）所说的“紧缩时代”。一些欧洲政府试图通过削减开支和提高税收来减少财政赤字。欧洲平均失业率从 2011 年的 10% 上升到了 2012 年的 11% 和 2013 年的 12%。就连频繁倡导紧缩的国际货币基金组织（IMF）也承认，欧洲紧缩措施的危害性大大超出了预期。

我们无法知道莱因哈特和罗戈夫的论文导致了多少产值损失，使多少人失去了工作。经济学家有时会进行这样一种游戏：指出自己希望成为哪一篇著名论文的作者。如果让我选择不希望由自己完成的论文，莱因哈特和罗戈夫的这篇论文将是我的首选之一。

事实证明，莱因哈特和罗戈夫的研究从一开始就存在严重缺陷。马萨诸塞大学安姆斯特分校研究生托马斯·赫恩登（Thomas Herndon）2012 年选修了一门由迈克尔·阿什（Michael Ash）和罗伯特·波林（Robert Pollin）教授的研究生统计课程，他的一项课程作业是复制一篇著名的研究论文。赫恩登选择了莱因哈特和罗戈夫的论文。他付出

了大量努力，但他无法复制他们的结果。他认为自己在某个地方出了差错，但他无法找到这个错误。

赫恩登的教授之前见过类似的情况，通常是因为学生出了差错。这也是该练习的目的之一——让学生学习仔细而正确地从事严肃的工作。这一次，情况有所不同。就连教授也无法发现这篇著名论文中的错误。

赫恩登认为知名教授不太可能花时间帮助一个小小的研究生，但他还是放弃了独自理清这项研究的努力，开始联系莱因哈特和罗戈夫。果然，他们对他最初的请求置之不理。不过，他一次又一次地发出请求。毕竟，要想通过这门课程，他必须弄清自己在哪里出了问题。

莱因哈特和罗戈夫最终向赫恩登提供了他们的数据以及计算时使用的电子表格。赫恩登很快发现了问题。实际上，这些数据存在一系列问题，包括粗心的错误和可疑的程序:

电子表格中的一个错误

20 个国家的原始数据以反向字母顺序出现在一张电子表格的 30 至 49 行。不过，当莱因哈特和罗戈夫编写带有电子表格编码的计算公式时，他们没有使用 30 至 49 行，而是使用了 30 至 44 行，因此他们忽略了五个国家（澳大利亚、奥地利、比利时、加拿大和丹麦）。其中，三个国家（奥地利、比利时和加拿大）拥有债务与 GDP 之比大于 90% 的时间段；在这些年份里，三个国家的增长率都是正值。这三个与论文结论相反的例子都被排除在了莱因哈特和罗戈夫的计算之外。

选择性忽略某些数据

在电子表格中，莱因哈特和罗戈夫的计算还忽略了其他一些数据。在许多情况下，这是因为一些国家在一些年份里缺少数据。不过，1946～1950 年的澳大利亚、1946～1950 年的加拿大以及 1946～949 年的新西兰是有数据的。奇怪的是，这些数据并没有被包含在计算之中。

新西兰的数据尤其重要，因为被忽略的四年是债务与 GDP 之比超过 90% 的五年中的四年（1951 年是第五年）。在这五年里，新西兰的 GDP 增长率分别是 7.7%、11.9%、-9.9%、10.8% 和-7.6%，平均增长率是 2.6%。莱因哈特和罗戈夫排除了前四年的数据，称新西兰高债务年份的平均增长率是-7.6%。

不同寻常的平均

莱因哈特和罗戈夫写道，表 4.1 中的计算基于“1186 个年度观测值，每个类别中包含大量观测值，包括 96 个超出 90% 的观测值。”你可能认为 90% 类别中的-0.1% 平均增长率是对 96 个年度观测值直接取平均得到的。事实并非如此。

莱因哈特和罗戈夫计算了每个国家的平均增长率，然后计算这些国家增长率的平均值。例如，英国有 19 年债务与 GDP 之比超过 90%，在这 19 年里，其平均 GDP 增长率为 2.4%；由于新西兰被忽略了四年，因此它只有一年债务与 GDP 之比超过 90%，在这一年，其 GDP 增长率是-7.6%。根据 20 个年度观测值，这两个国家的平均增长率是 1.9%。不过，莱因哈特和罗戈夫计算了 2.6% 和-7.6% 的平均值，得到了-2.5% 的平均增长率。

我们不知道莱因哈特和罗戈夫是无意中犯了一个错误，还是故意选择了这种不同寻常的计算方法，为一个国家一年的数据和另一个国家 19 年的数据赋予了相同的权重。不过，我们至少知道这个错误支持了他们的观点。

表 4.2 显示了这些问题的效果。10 个国家经历过债务与 GDP 之比超过 90% 的年份。在 9 个国家，平均增长率是正值。唯一的例外是战后的美国，而且这个例外与研究的主题无关。

表格的前两列显示了实际高债务年份以及莱因哈特和罗戈夫在计算中包含的年份。莱因哈特和罗戈夫忽略了前三个国家以及新西兰四个年份的数据。所有这些被忽略数据的平均增长率都是正值。

后面两列显示了平均增长率，其中“正确”一列包括了被莱因哈特和罗戈夫忽略的三个国家以及新西兰的四个年份。

表 4.2：莱因哈特和罗戈夫（RR）忽略的结果

	年份数量		GDP 增长	
	正确	RR	正确	RR
澳大利亚 1946～1950	5	0	3.8	
比利时 1947，1984～2005，2008～2009	25	0	2.6	
加拿大 1946～1950	5	0	3.0	
希腊 1991～2009	19	19	2.9	2.9
爱尔兰 1983～1989	7	7	2.4	2.4
意大利 1993～2001，2009	10	10	1.0	1.0
日本 1999～2009	11	11	0.7	0.7
新西兰 1946～1949，1951	5	1	2.6	-7.6
英国 1946～1964	19	19	2.4	2.4

美国 1946～1949	4	4	-2.0	-2.0

总体而言，通过忽略数据以及为所有国家赋予相同的权重，莱因哈特和罗戈夫计算出的高债务年份平均 GDP 增长率为-0.1%。如果将失踪的数据包含进来，并且考虑到高债务年份的数量，正确的平均值为 2.2%。赫恩登、阿什和波林的结论是："与'莱因哈特和罗戈夫'的观点相反，公共债务与 GDP 之比超过 90% 时的平均 GDP 增长率与公共债务与 GDP 之比较低时的平均 GDP 增长率并没有明显的区别。"

表 4.1 所显示的莱因哈特和罗戈夫的计算结果讲述了一个清晰而令人信服的故事：过高的债务将不可避免地降低增长速度；如果超过 90% 的临界点，国家将会陷入衰退。除了计算中的各种错误，这些总结性的统计量还隐藏了数据中存在的巨大波动性。债务与增长之间并不存在简单而紧密的关系。

图 4.1 显示了论文所依据的数据，图中 90% 的阈值处有一条竖线；根据研究的观点，超过这条竖线的高债务将使一个国家陷入衰退。我们可以看到，债务与增长之间并没有任何令人信服的关系。许多高债务年份也是高增长年份。许多低债务年份也是低增长年份。这些数据本身证明了 90% 这一比例没有任何特殊之处。

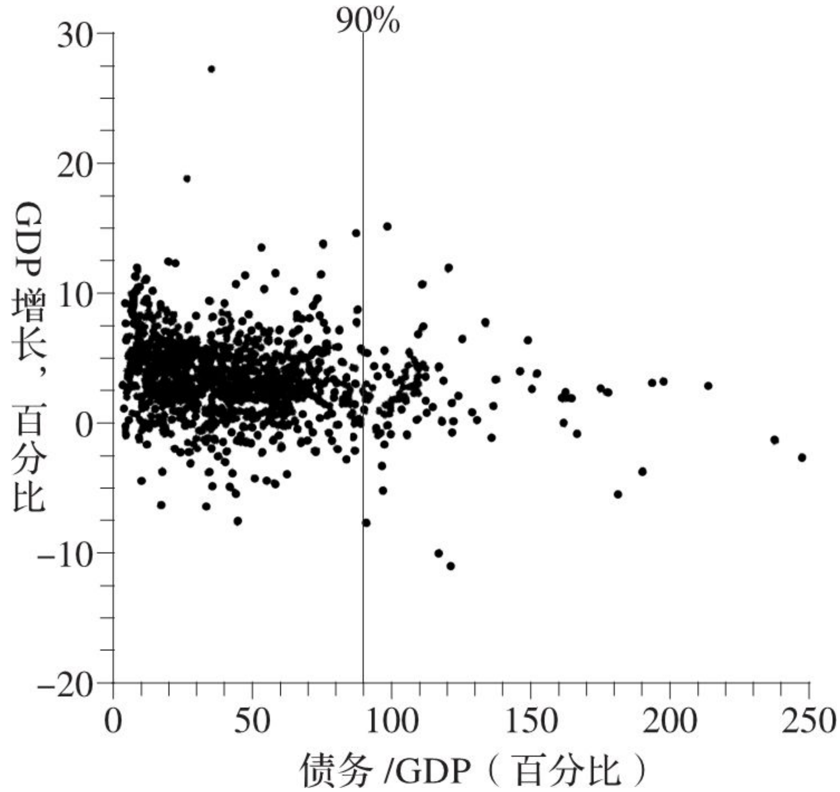


图 4.1　债务等同于衰退?

相关性等同于因果关系?

财政紧缩的支持者认为，莱因哈特和罗戈夫的研究不仅证明了提高政府债务将会降低经济增长速度，而且提出了一个严厉的警告：如果政府债务水平超过 GDP 的 90%，可能会导致经济衰退。

也许，这种因果关系应该倒过来：经济衰退将导致债务比率上升。首先，有一个简单的算术性观点：GDP 的下降将直接提高债务与 GDP 的比率。假设债务是 85，GDP 是 100，那么债务与 GDP 之比为 85%：

债务 / GDP = 85 / 100 = 0.85

如果 GDP 下降到 90，债务与 GDP 之比将提升至 94%：

$$债务 / GDP = 85 / 90 = 0.94$$

不是债务与 GDP 之比上升导致 GDP 下降，而是 GDP 下降导致债务与 GDP 之比上升。

不过，这还没完。在经济衰退期间，政府税收将会下降，政府在失业保险、食品券以及其他安全保障上的开支将会增长。这两个因素都会推高政府债务。因此，经济衰退不仅会使债务与 GDP 之比的分母变小，而且会使分子变大。

如果经济增长与债务/GDP 之间存在统计相关性，这种相关性可能主要（甚至完全）来自经济对债务比率的影响，而不是反方向的影响。不是高债务比率降低了增长速度，而是增长速度的降低提高了债务比率。

马萨诸塞大学安姆斯特分校的另一位教授阿林德拉吉特·杜比（Arindrajit Dube）也对这个问题进行了研究。根据莱因哈特-罗戈夫的数据，杜比发现，债务与 GDP 之比与过去经济增长之间的相关性高于它与未来经济增长之间的相关性——这恰恰印证了“经济增长导致债务与 GDP 的比率发生变化”这一观点。

堕胎会减少犯罪吗?

30 年前，数学理论家是经济学的上帝。对现实世界一无所知几乎是一种荣誉的象征。当杰拉德·德布鲁（Gerard Debreu）1983 年获得诺贝尔奖时，记者想让他对罗纳德·里根（Ronald Reagan）的经济政策说点什么。德布鲁坚决拒绝发表任何言论。一些人怀疑他不知道或者不关心这种事情。

时代发生了改变。经济学理论家由于脱离现实的假设和明显的错误结论而受到了攻击。与数据打交道的经验主义者成了新的经济学上帝。这个时代的标志之一是芝加哥经济学家史蒂文·莱维特（Steven Levitt）和《纽约时报》记者史蒂文·都伯纳（Steven Dubner）合著的畅销书《魔鬼经济学：一位草莽经济学家探索一切事物不为人知的一面》。

书名的选择完全是为了增加销量，具有很大的误导性。这本书与魔鬼没有任何关系，莱维特也不是草莽经济学家。而且，这本书并没有探索一切事物不为人知的一面。实际上，莱维特是典型的正统经济学家。他在哈佛大学获得了学士学位，在麻省理工学院获得了博士学位。他目前是芝加哥大学教授。2003 年，美国经济学会向莱维特颁发了约翰·贝茨·克拉克奖章。这个奖章用于奖励美国四十岁以下最优秀的经济学家，而且常常是诺贝尔奖的前瞻。莱维特是经济学的经验主义上帝之一。

莱维特善于以仔细而巧妙的方法利用数据解决有趣的问题，比如竞选经费对选举结果的影响，监禁对犯罪率的影响，教育激励对教师作弊行为的影响。他的结论往往会引起争议：竞选经费的规模并没有太大的影响；每个由于诉讼案件过多而被放走的囚犯都会导致每年增加 15 起犯罪案件。根据学生的考试成绩奖励教师的做法会极大地增加由教师辅助的作弊现象。莱维特曾写道："我们最喜欢做的事情就是在数据中发现其他人看不到的东西。"

不过，上帝也会犯错误。

莱维特与约翰·多诺霍（John Donohue）合写的一篇论文认为，美国的合法堕胎降低了总体犯罪率。这篇论文也许是莱维特最著名的论文。文章指出，如果没有合法堕胎，那么由于社会经济环境以及/或者家长的忽视，那些"没有必要"但仍然被生下来的孩子将会产生犯罪倾向（尤其是暴力犯罪）。

这是一种有趣的理论，甚至有可能是正确的。问题是，由于存在众多的遗传和环境因素，因此我们很难弄清每个因素的重要性。莱维

特说，他喜欢将结果从数据中梳理出来。听起来，他似乎在炫耀一种宝贵的技能。不过，这似乎也意味着他喜欢搜刮数据，而这可能是一种危险的做法。

下面是莱维特所依据的数据。1970 年，堕胎在美国五个州实现合法化。1973 年 1 月 22 日，由于“罗诉韦德案”，堕胎在全国范围内实现合法化。图 4.2 显示，在“罗诉韦德案”18 年以后，美国的谋杀率在 1991 年达到峰值。随后，谋杀率开始下降，这意味着堕胎的确导致了犯罪人口的减少。在考察每个州的各项数据以后，多诺霍和莱维特总结道：“对于最近犯罪现象的减少，合法堕胎似乎起到了 50% 的作用。”

这种争议性说法受到了来自各个角度的批评。一些人认为，谋杀率之所以在 20 世纪 80 年代上升并在 90 年代下降，是因为敌对贩卖可卡因团伙之间的地盘争夺战出现了起伏。另一种理论是，20 世纪 90 年代的犯罪率之所以下降，是因为孩子们由于接触含铅汽油和油漆而中毒的机会变少了。

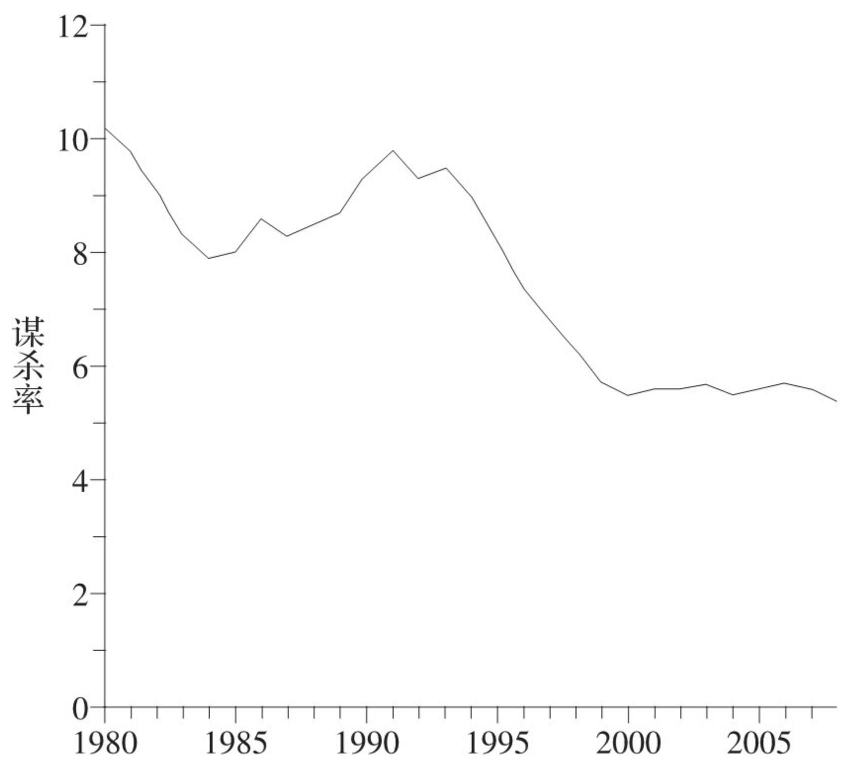


图 4.2　每 10 万人的谋杀案数量

此外，一些研究人员指出，对于自身行为对未来造成的后果有着清醒认识的女性可能更倾向于选择堕胎，而她们也是能够培养出守法孩子的优秀母亲。而且，合法堕胎的规定可能会增加性活动和非婚生子女，而在单亲家庭里长大的孩子可能更容易犯罪。这些观点意味着合法堕胎反而有可能提高犯罪率。

让我们观察一些数据。图 4.3 显示了 14 ~ 17 岁的犯罪率。“罗诉韦德案”所导致的首批合法堕胎行为扼杀了将在 1973 年年末出生的孩子。这些没有必要出生的孩子将在 1987 年年末达到 14 岁，并将在 1991 年年末达到 18 岁。因此，首批受到“罗诉韦德案”影响的群体将在 1987 年年末和 1991 年年末之间达到 14 ~ 17 岁。图 4.3 显示，犯罪率在这些年份持续上升，并在 1993 年达到峰值。1993 年的犯罪率是 1984 年的三倍。不过，1993 年的那批人出生时堕胎是合法而方便的，而 1984 年的那批人出生时堕胎是非法而困难的。



图 4.3　14 ～ 17 岁的谋杀率

图 4.4 显示了 18 ~ 24 岁以及 25 ~ 34 岁的犯罪率。首批受到“罗诉韦德案”影响的 18 ~ 24 岁群体的谋杀率出现了下降，但首批受到“罗诉韦德案”影响的 25 ~ 34 岁群体的谋杀率出现了上升。

所以，“罗诉韦德案”后 18 ~ 24 岁群体的谋杀率出现了下降，但 14 ~ 17 岁和 25 ~ 34 岁群体的谋杀率出现了上升。这些数据与“罗诉韦德案”降低了谋杀率的理论并不一致。图 4.3 和图 4.4 真正能够证明的是谋杀率在 20 世纪 90 年代早期达到峰值：14 ~ 17 岁群体以及 18 ~ 24 岁群体为 1993 年，25 ~ 34 岁群体为 1991 年。35 岁以上群体（图中没有显示）的谋杀率在 1990 年达到峰值。换一种说法，35 岁以上群体的谋杀率在 1990 年达到峰值，25 ~ 34 岁群体在 1991 年达到峰值，这些人都是在“罗诉韦德案”之前出生的。18 ~ 24 岁群体的谋杀率在 1993 年达到峰值，其中一些人在“罗诉韦德案”之前出生，一些人在其之后出生。最后，14 ~ 17 岁群体的谋杀率在 1993 年达到峰值，这些人都是在“罗诉韦德案”发生几年以后出生的。

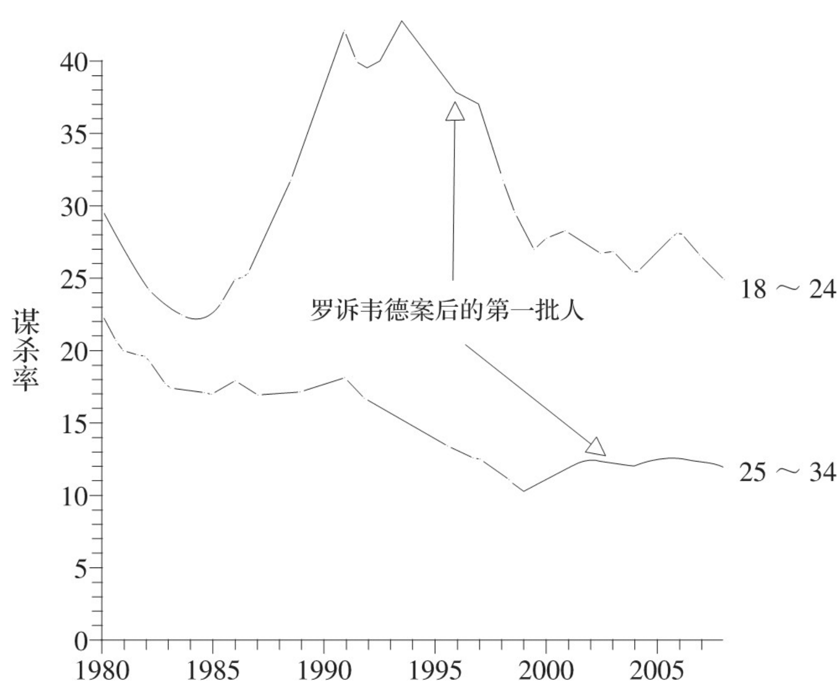


图 4.4　18 ~ 24 岁以及 25 ~ 34 岁的谋杀率

真相是，所有年龄群体的谋杀率都在 20 世纪 90 年代早期达到峰值。不管怎样解释，合法堕胎都不会是其中的原因。

是我的错

那么，多诺霍和莱维特是如何得出“合法堕胎减少犯罪”这一结论的呢？波士顿联邦储备银行的两位经济学家克里斯多夫·富特（Christopher Foote）和克里斯多夫·戈茨（Christopher Goetz）对多诺霍和莱维特的研究进行了仔细检查，发现了三个问题。首先，多诺霍和莱维特使用了逮捕日期，但是逮捕与犯罪之间可能间隔了一年以上的时间。其次，多诺霍和莱维特考虑的是犯罪总数量，而不是犯罪率（调整了人口规模因素后的犯罪数量）。较少的人口自然可能犯下较少的罪行，而合法堕胎本身就会降低人口数量。

第三个问题是“程序错误”，也就是一个简单的人为错误。多诺霍和莱维特声称，他们利用某种统计程序处理各州内部不同年份之间的变化，但他们并没有这样做。在修正这些错误以后，富特和戈茨得到的结论是，堕胎很可能会提高暴力犯罪率，同时对财产犯罪率没有影响。

莱维特承认了这个程序错误，并表示：“这使我个人感到非常尴尬，因为我一直为自己对数据的仔细处理而感到自豪。”不过，这不是莱维特第一次感到如此尴尬。之前一篇关于“提高警力对犯罪的影响”的论文也出现了一个程序错误，对此他发表了这样的致歉声明：“我需要承认这些错误，这使我个人感到极为尴尬。”在这两次事故中，莱维特很有风度地承认了错误（他还有别的选择吗？），但他

仍然认为其他证据可以支持他的结论，只是结论的效果可能比他最初预想的要小一些。也许他说得没错，但是人们很难忘记这些尴尬的时刻。

×如何轻松识破一本正经地胡说八道

具有争议性的论断之所以具有争议性，是因为它们违反直觉——这是一个很好的怀疑理由。当你听到这样的说法时，不要轻易认为自己是错误的。具有争议性的论断很可能应当被丢弃。考虑数据是否存在问题，比如自选择偏差。考虑因果关系是否应该颠倒过来。哦，对了，还要考虑出现错误的可能性（比如人们让计算机计算 196 的平方根，而不是 169 的平方根。即使是最优秀、最诚实的研究人员也是人），而人总会犯错误。

莱因哈特和罗戈夫具有全球影响力的研究以及《魔鬼经济学》的巨大成功给我们带来的具有讽刺意义的教训是，数据并不比思想更加重要。我们常常会被数据欺骗。

第 5 章
扬基队的门票真的划算吗?

图像可以帮助我们解释数据，做出推断，发现倾向、模式、趋势和关系。一张图片的价值不仅有可能胜过千言万语，而且有可能胜过一千个数字。不过，从本质上说，图像具有描述性——图片的目的是讲故事。和其他故事一样，愚笨的人可能会毁掉一个包袱，不诚实的人可能会撒谎。

图像可能会有意无意地扭曲或破坏数据。

我的天哪

在一次例行会议上，一家互联网公司的分析团队向总裁展示了图 5.1 中的收入图像。图中的数据是公司过去 7 个季度的收入。老实说，这张图非常无聊。

总裁看了一会儿，说：“那么，为什么这张图的收入这么平坦，而我上周向董事会展示的那张图的收入那么糟糕呢？”分析团队感到很吃惊。他们并没有见过另一张图，只好耸了耸肩。显然，他们无法回答这个问题。

总裁将制作上一张图的那个财务人员叫了过来，让他把那张图（图 5.2）分发给大家。当分析人员看到这张图时，他们立刻笑了起来，因为这张图的纵轴上并没有零点。

总裁没有笑。她说，董事会一直在盘问她，让她解释为何收入下降得如此剧烈。她一次又一次地辩解说："不是这样的！"董事会成员则一次又一次地将手指向这张显示收入崩塌的该死图表。

图 5.1 和图 5.2 使用了完全相同的数据。不过，图 5.1 的纵轴包含零点，图 5.2 的纵轴则不含零点。图 5.1 掩盖了数据中的任何模式，我们无法判断利润达到峰值的时间。另一方面，图 5.1 正确地向我们传达了"最近的下探很轻微"这一信息。

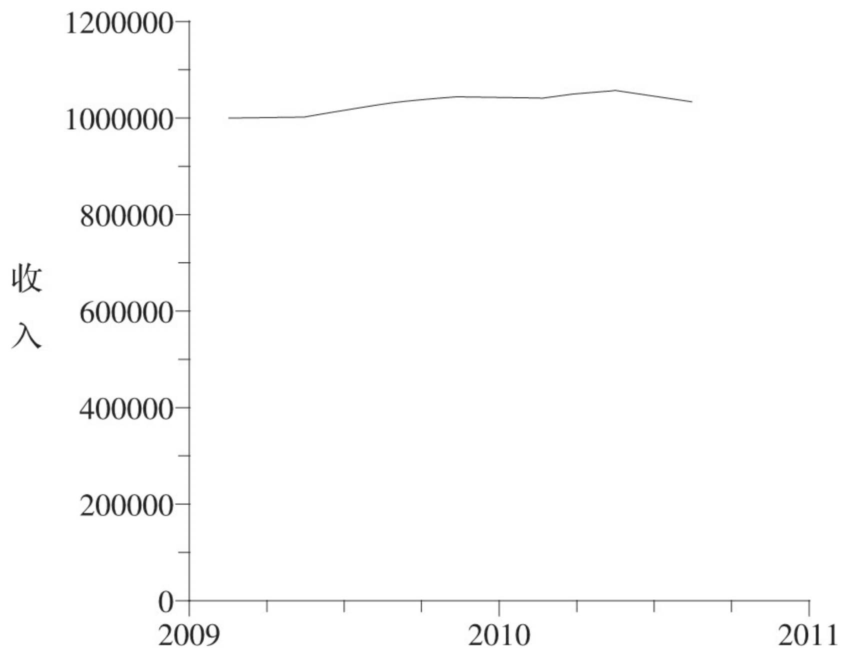


图 5.1　收入是平的

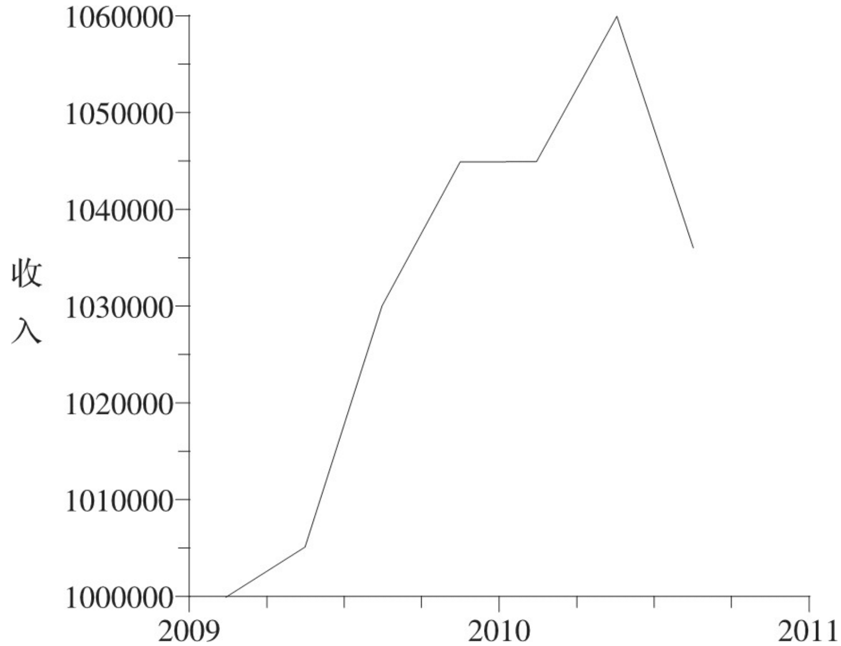


图 5.2　收入急剧下降

图 5.2 忽略了零点，放大了数据的波动性，使我们能够发现原本非常模糊的变化。不过，当零点被忽略时，图像将不再准确指示变化的量级。要想知道变化幅度，我们需要查看具体数据。图 5.2 中的曲线高度下降了 40%，但实际收入只下降了 2%。

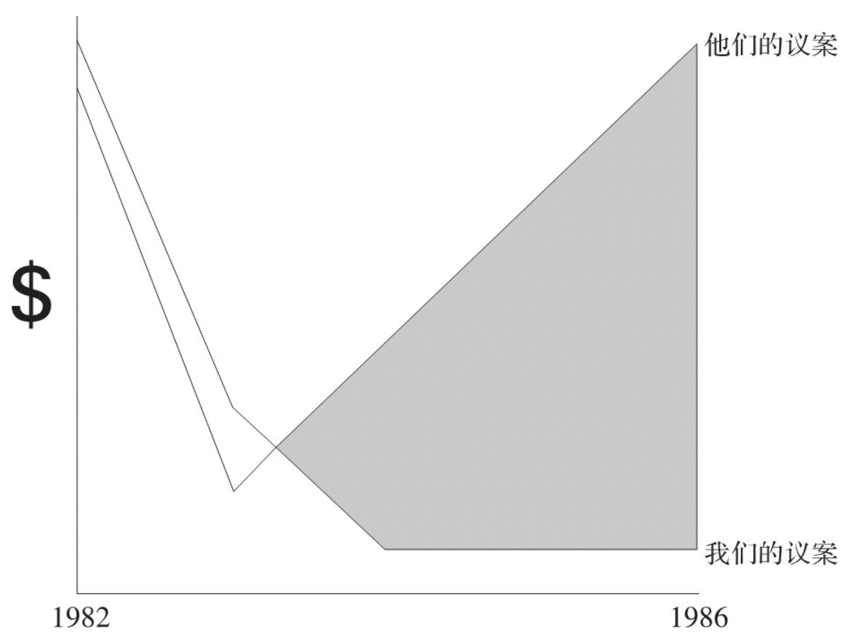


图 5.3　神秘的数轴

图 5.3 显示的图像具有更低的透明度，因为纵轴上并没有数字。罗纳德·里根总统曾在电视上用这样的图像对年收入 2 万美元的家庭根据他的议案以及众议院筹款委员会拟定的计划分别需要支付的税金进行比较。纵轴显然忽略了零点，而且并没有显示任何数字，只有一个巨大的美元符号。由于纵轴没有数字，因此我们无法衡量两条线之间差距的幅度。实际上，1986 年的差距相当于 9% 的税收减少，即从 2385 美元下降到 2168 美元。由于纵轴忽略了零点，因此这个 9% 的差距被放大成了标有“他们的议案”那条曲线高度的 90%。又由于纵轴上没有数字，因此读者无法发觉这种具有误导性的放大。

事后，白宫发言人戴维·格根（David Gergen）告诉记者：“我们尝试了带有数字的图像，发现它们在电视上很难看清，因此我们去掉了数字。我们只是想要传达一种思想。”是的，没错。

虚惊一场

1976 年，美国国家科学基金会（NSF）制作了一张图（图 5.4）。显然，这张图意味着颁发给美国公民的诺贝尔科学奖（化学、物理、医学）的数量出现了令人震惊的下降。我们的教育系统出了什么问题？我们的人才是否正在枯竭？

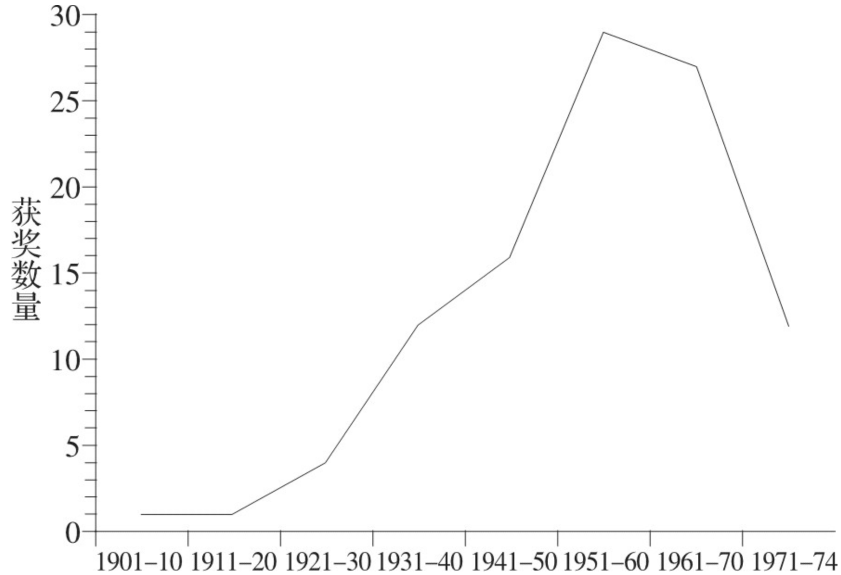


图 5.4　美国诺贝尔科学奖数量的急剧下降?

等一下。看看时间轴。前七个时间段都是 10 年，但是第八个时间段 1971 ~ 1974 年只有 4 年。由于四年颁发的诺贝尔奖数量少于 10 年颁发的数量，因此美国国家科学基金会制造了一种幻觉，使人误以为美国人获得的科学奖数量出现了下降。

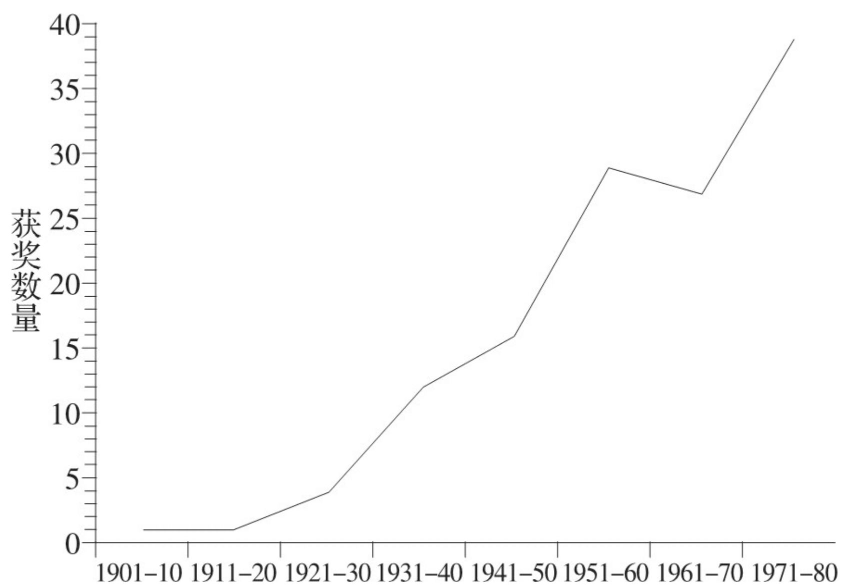


图 5.5　美国诺贝尔科学奖持续增加

图 5.5 使用了完整的 10 年期数据。如图所示，美国公民最终在 20 世纪 70 年代获得了比 60 年代更多的诺贝尔奖。当然，这个趋势无法永远持续下去，除非诺贝尔奖的总数量得到提升。不过，在 20 世纪 70 年代，美国获得了超过一半的诺贝尔科学奖，优于 20 世纪 80 年代、90 年代以及 21 世纪前 10 年的表现。

让他们吃蛋糕吧[①]

新保守主义者戴维·弗鲁姆（David Frum）在《纽约时报》上发表了一篇文章《欢迎，新富人》。图 5.6 是文章中一张图片的更新版。如图所示，在 1980 年到 1990 年之间，年收入超过 10 万美元的家庭数量出现了急剧增长。弗鲁姆写道："这个星球的历史上从未出现过如此庞大的富人群体。"这听上去是不是有点夸张？

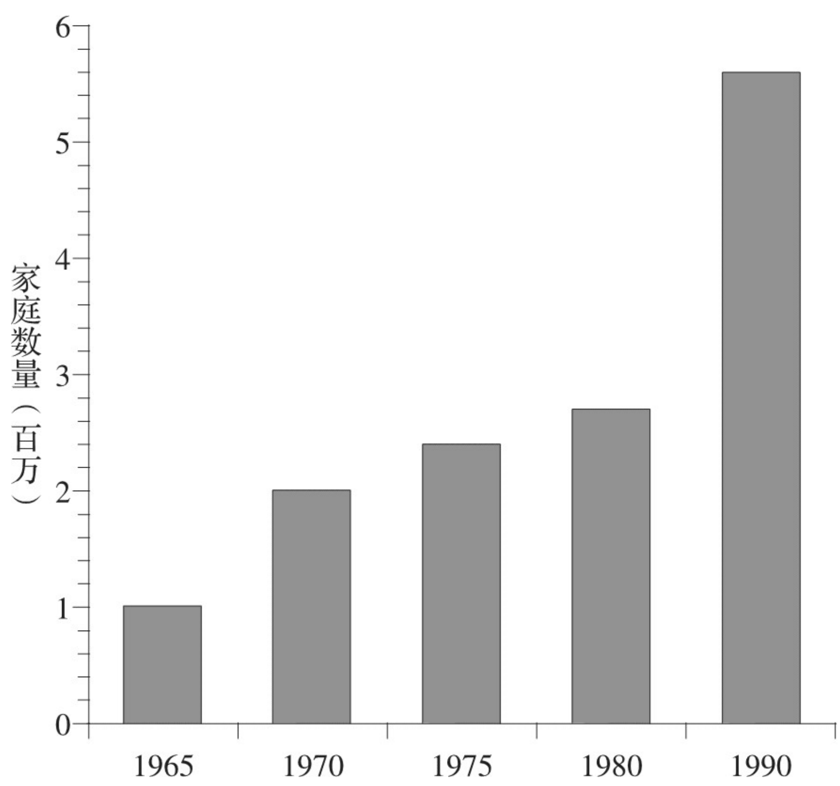


图 5.6　年收入超过 10 万美元的家庭数量

你注意到这张图的奇怪之处了吗？前四个条柱之间都是相隔五年，但是第四个条柱和第五个条柱之间却隔了 10 年（1980 和 1990）。如果调整条柱之间的间隔，插入 1985 年的条柱，这种增长就会变得更加平缓，不会在 1980 年和 1990 年之间出现突兀的上涨。

而且，1990 年的 10 万美元并不等同于 1965 年的 10 万美元。在这段时间里，物价上涨了大约三倍，因此 1990 年的 10 万美元大概相当于 1965 年的 2.5 万美元。考虑到通胀因素，我们应当对 1965 年年收入 2.5 万美元的家庭数量与 1990 年年收入 10 万美元的家庭数量进行比较。我们还应当考虑到 1965 年和 1990 年之间的人口增长。当总人口增加时，拥有高收入的人口数量自然有可能增加。

图 5.7 修正了所有这些问题，显示了扣除通胀因素后年收入超过 10 万美元的家庭百分比，并且插入了 1985 年的数据。此外，图中还包含了 1995 年和 2000 年的数据，以提供更多历史背景。根据调整后的数据，20 世纪 80 年代没有什么奇特之处。真正值得注意的是互联网泡沫所在的 90 年代后期。

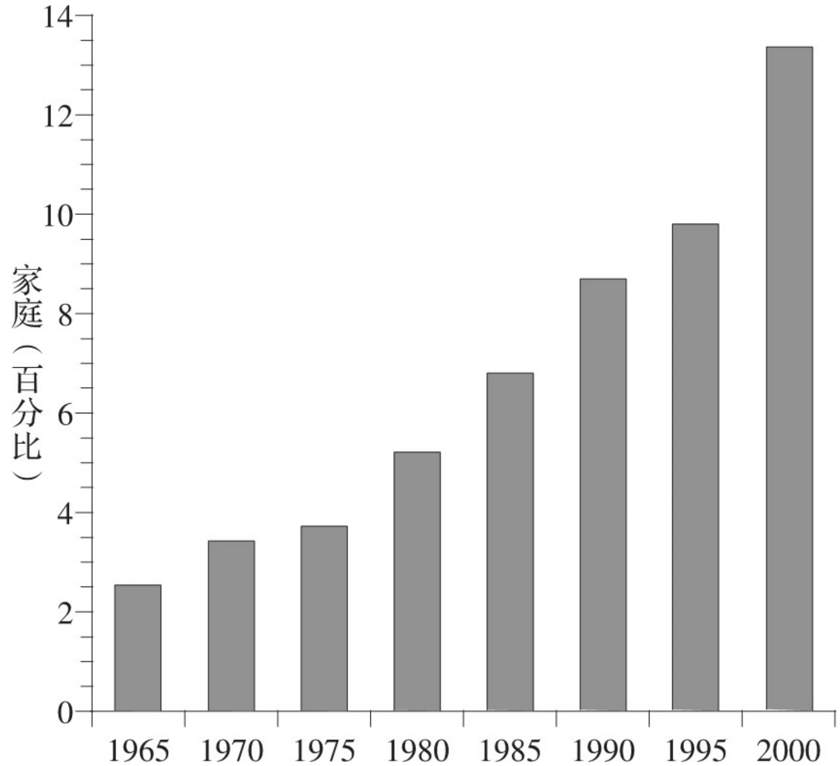


图 5.7　扣除通胀因素后年收入超过 10 万美元的家庭百分比

有弹性的数轴

图 5.8 是《华盛顿邮报》一张图表的更新版本。根据这张图，美国医疗成本从 1940 年到 2010 年出现了稳定的增长。（1940 年的成本被调整成了单位 1。）

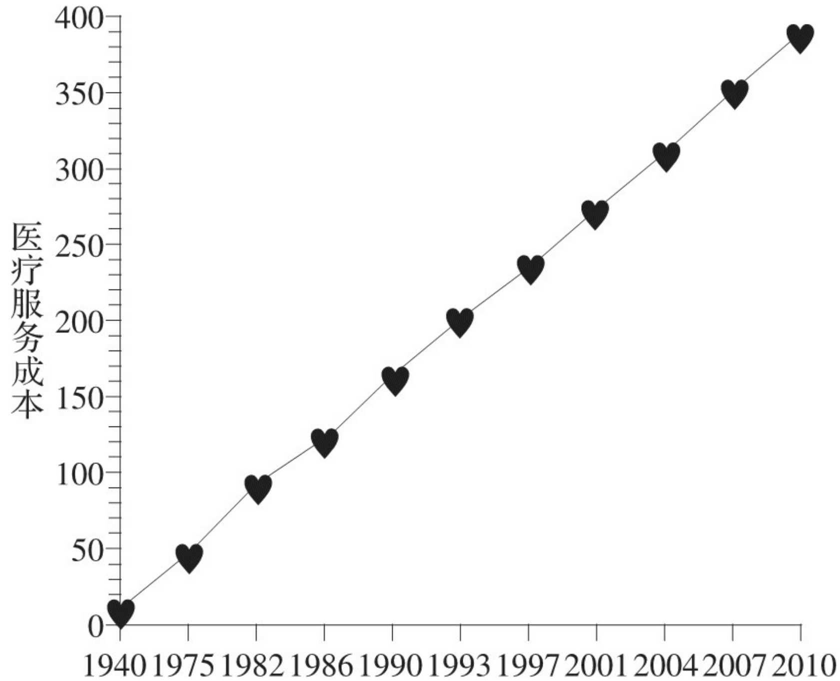


图 5.8　医疗成本稳步上升

这张直线图上装饰着心形图案。也许，这是为了显示医生的确非常关心病人，或者暗示丘比特那支笔直的箭射穿了十一颗心。或者，这种毫无必要的装饰也许是为了转移人们对横轴的注意力。在横轴上的十个等距区间中，第一个区间表示 1940 年到 1975 年之间的三十五年，第二个区间表示 1975 年到 1982 年之间的七年，其余区间表示三年或者四年。这些令人困惑的区间显然是为了使曲线看上去更加平直。直线在视觉上也许更具吸引力，但它会歪曲数据。要想让图像准确显示数据，每条数轴上的单位必须具有一致性。同样的 1 厘米不能时而代表 35 年，时而代表七年，时而代表三年。如果图中的区间变来变去，数据一定会遭到扭曲。

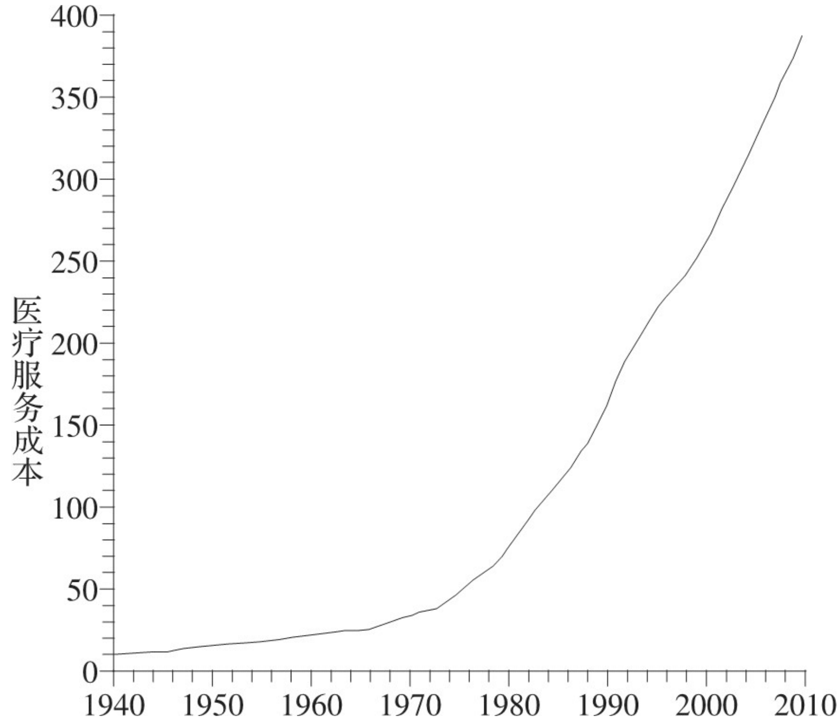


图 5.9　更好的医疗成本图

图 5.9 使用了一致的区间。如图所示，医疗成本并没有沿着直线增长。

为什么成本在 1970 年以后加速增长？也许是因为始于 1965 年的医疗保险开始增长。也许是因为价格整体上出现了增长（这些图像应当扣除医疗成本的通胀因素）。我们无法根据这些数据本身做出判断。不过，我们至少知道，图表中间隔的变化会扭曲数据。

将恶作剧翻倍

1786 年，拥有多项才能的苏格兰人威廉·普莱费尔（William Playfair）出版了包含 44 张图表的《商业和政治图表集》，这是历史上用图像表示数据的一个标志性事件。“普莱费尔”的字面意义是“公平

行事”，尽管这个人并不总是能够诚实地运用自己的才能。在这份图表集中，有 43 张图表显示了工资、物价以及其他数据随时间的变化。这种目前已经司空见惯的工具在当时是一件极为新奇的事物，因此普莱费尔不得不做出解释：“这种方法使一些人感到震惊，他们觉得这是一种荒谬的做法，因为几何测量与金钱和时间之间没有任何关系；但是这里却用几何工具来表示金钱和时间。”

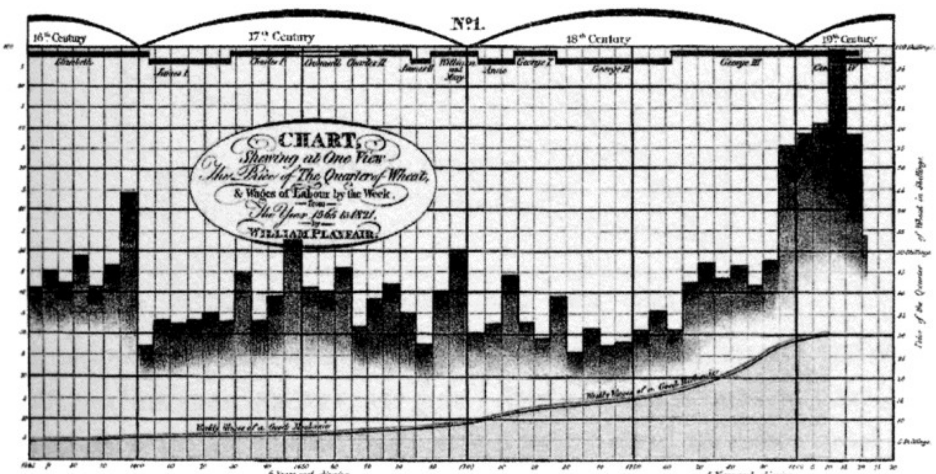


图 5.10　普莱费尔跨越 250 年的工资和物价图像

普莱费尔最著名的一张图发表于 1821 年，显示了从 1565 年到 1821 年“优秀技工”的周薪（图表下方的曲线）和小麦价格（黑色条柱）。普莱费尔写道：“值得注意的最主要的现象是，历史上，小麦相对于技工劳动的价格从未像现在这样便宜。”虽然普莱费尔得出了正确的结论，即工资增长超过了小麦价格的增长，但他的图表似乎显示出了相反的结论。

我将使用 1975 年到 2010 年美国家庭收入中值与消费者物价指数（CPI）的数据来说明普莱费尔的图表为何具有误导性。将这些数据放在一张图表上是一件很有挑战性的工作，因为收入的范围是 1.2 万到 5 万美元，而消费者物价指数的范围则是 54 到 218。一种解决方案（也就是普莱费尔的解决方案）是使用两个纵轴，一个表示收入，另

一个表示消费者物价指数。图 5.11 似乎清晰表明，家庭收入的增长远远超出了消费者物价指数的增长。现在的美国人在经济上显然比过去富裕得多。

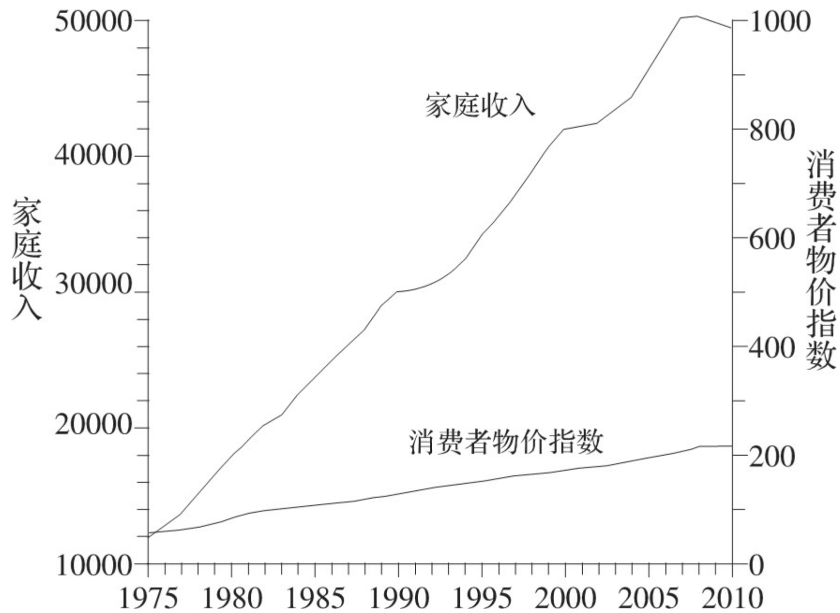


图 5.11　收入增长远超物价

图 5.11 真正说明的问题是，如果你将数轴放大一倍，你就可以将恶作剧放大一倍。通过使用两个纵轴并忽略一个或两个纵轴的零点，你可以创建一家“统计美容院”，开启许多美容可能性。图 5.12 显示了完全相同的数据，但是图中物价的增长似乎远远超过了收入的增长！这个把戏的关键在于调整数轴上显示的数字。如果在数轴上选择范围较小的数字，数据的起伏就会被放大，如果在数轴上选择范围较大的数字，数据的起伏就会被抚平。图 5.11 使用了范围较小的收入数轴和范围较大的 CPI 数轴，因此收入的增长看上去超过了物价的增长。图 5.12 采用了相反的做法，使用了范围较大的收入数轴和范围较小的 CPI 数轴，因此物价的增长看上去超过了收入的增长。

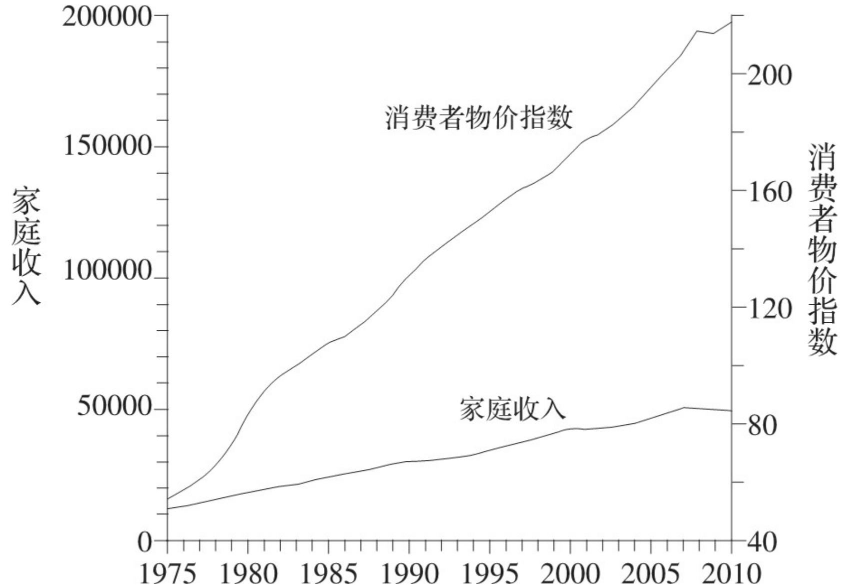


图 5.12　物价增长远超收入

这个问题没有适用于所有情况的解决方案。在上面的例子中，一个很好的办法是用物价数据校正收入数据，得到一组扣除通胀因素后的收入数据。扣除通胀因素后，过去 35 年的家庭收入中值出现了一定的增长（收入增长速度稍快于物价），但是这个过程中也存在一些由于经济衰退而导致的下挫。

学生谎言

一所中等规模的大学有一份校刊，叫做《学生生活》。在校园里，人们亲切地称之为“学生谎言”。这份校刊制作了一张图（图 5.13）。在这张图上，我们可以明显看出，在 20 世纪 90 年代，该校的就读成本（学费加上住宿和餐饮费用）迅速上涨，而学校在《美国新闻》上的排名却急剧下降。

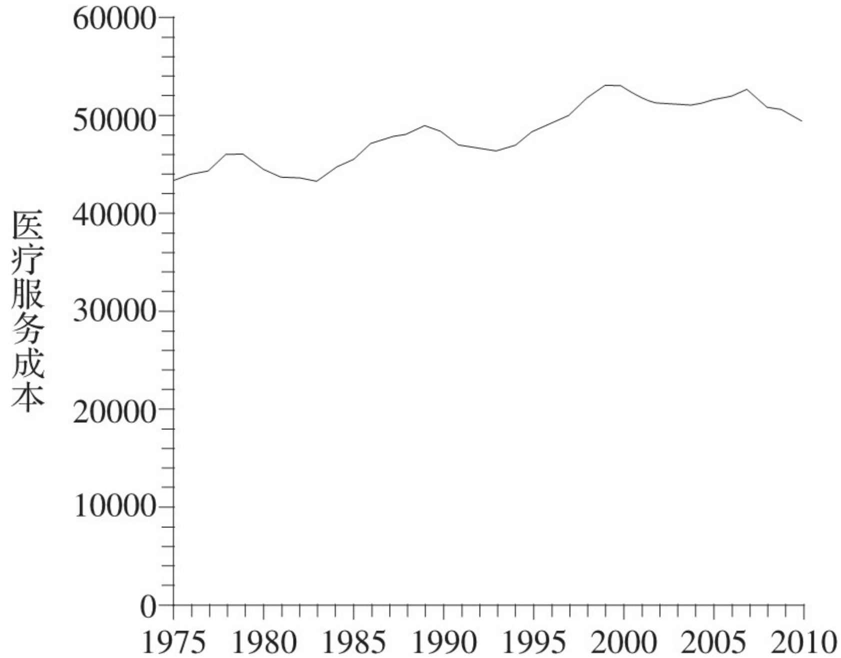


图 5.13　扣除通胀因素后的收入

图中的问题实在是太多了。从哪儿说起呢？图 5.14 有两个纵轴，一个用于表示成本，一个用于表示排名。两个纵轴都忽略了零点，以放大成本和排名的变化。不仅零点被省略，整个数轴和相关数字都被擦掉了，因此我们无法衡量 20 世纪 90 年代这些变化的大小。这张图所依据的数据显示，成本从 1990 年的 2 万美元增长到了 2000 年的 3.6 万美元，每年大约增长 6%，这个数字不算小，但它与同类大学的增长速度相当。

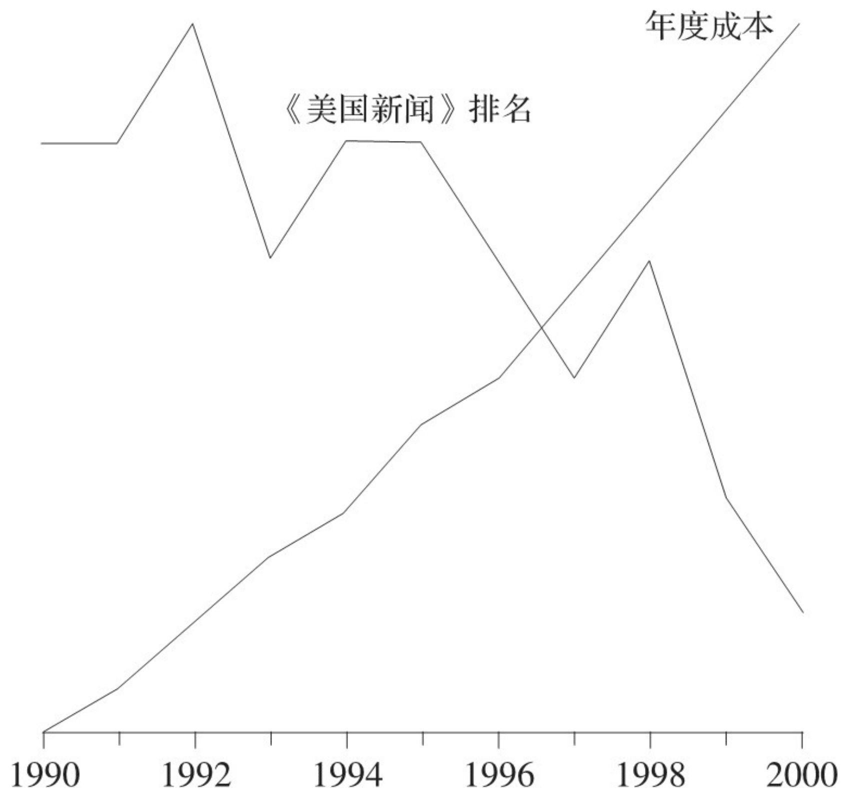


图 5.14　花费越多，收获越少

《美国新闻》的排名从 17 变成了 13，这是一种很大的变化。不过，等一下。在大学排名中，第一名才是最好的学校！从 17 变成 13 实际上是一种进步。

因此，成本的年增长率为 6%，与同类大学基本持平，《美国新闻》排名则出现了很大的进步，这与图 5.14 讲述的令人震惊的故事完全不同。在此之前，《伊萨卡时报》头版曾经贴出一张关于康奈尔大学学费和排名的类似图表，它比图 5.14 更加复杂，而且更加古怪。除了图 5.14 中的问题，《伊萨卡时报》的图表还省略了横轴，因为它比较的是康奈尔大学在两个不同时间段的学费和排名！学费使用的是 1965 ~ 1999 年（35 年）的数据，学校排名使用的则是 1988 ~ 1999 年（12 年）的数据。

扬基队的门票很划算

纽约扬基队是棒球领域最具传奇色彩的球队。截止 2013 年，在 108 次世界职业棒球大赛中，扬基队获得了 40 次参赛资格，赢得了 27 次冠军。排名第二的圣路易斯红雀队只获得了 11 次冠军，被远远甩在了后面。43 名扬基队队员进入了名人堂，包括贝比·鲁斯（Babe Ruth）、卢·贾里格（Lou Gehrig）、乔·狄马乔（Joe DiMaggio）和米基·曼特尔（Mickey Mantle）。扬基队也是职业体育界最受人鄙视的球队之一，部分原因在于，人们认为这支球队利用高薪挖走了小城市不太富裕的球员，通过金钱手段获得了成功。一些棒球迷自豪地戴着扬基棒球队的帽子，其他一些棒球迷则故意炫耀 T 恤上的“扬基糟糕透顶”字样。

扬基体育场 1923 年启用，其成本相当于今天的 3000 多万美元，拥有前所未有的 5.8 万个席位。贝比·鲁斯是 20 世纪 20 年代最受欢迎的棒球选手，他在扬基体育场的首场比赛中击出了全垒打。观看鲁斯全垒打的观众购买的门票被用于支付体育场的开销，这使它获得了“鲁斯建造的建筑”这一外号。86 年后，2009 年，扬基队搬到了街对面新的扬基体育场，该体育场的建设成本超过 20 亿美元。在新体育场的建造过程中，许多球迷担心扬基队会逐步提高已经很贵的票价，以支付体育场的费用。

扬基队用体育场包厢座位价格的历史数据制作了一张图（图 5.14）。这张图显示，从 1995 年到 2010 年，门票价格并没有加速上涨，反而趋于平缓，这与人们的感受相反。

你注意到图 5.15 的异常之处了吗？不知为什么，图中的时间被放在了纵轴上。我们平时习惯于看到时间出现在横轴上的图像，因此我们很难第一眼看清图 5.14 的含义。图 5.16 显示了同样的数据，只是将时间放在了正常的位置，即横轴上。这种数轴转换使结论发生了逆转。1994 年以后，扬基队的票价增速不仅没有下降，反而出现了上

升。1967 年到 1994 年的年增长率为 6%，1994 年到 2010 年的年增长率为 21%。你还觉得这很划算吗?

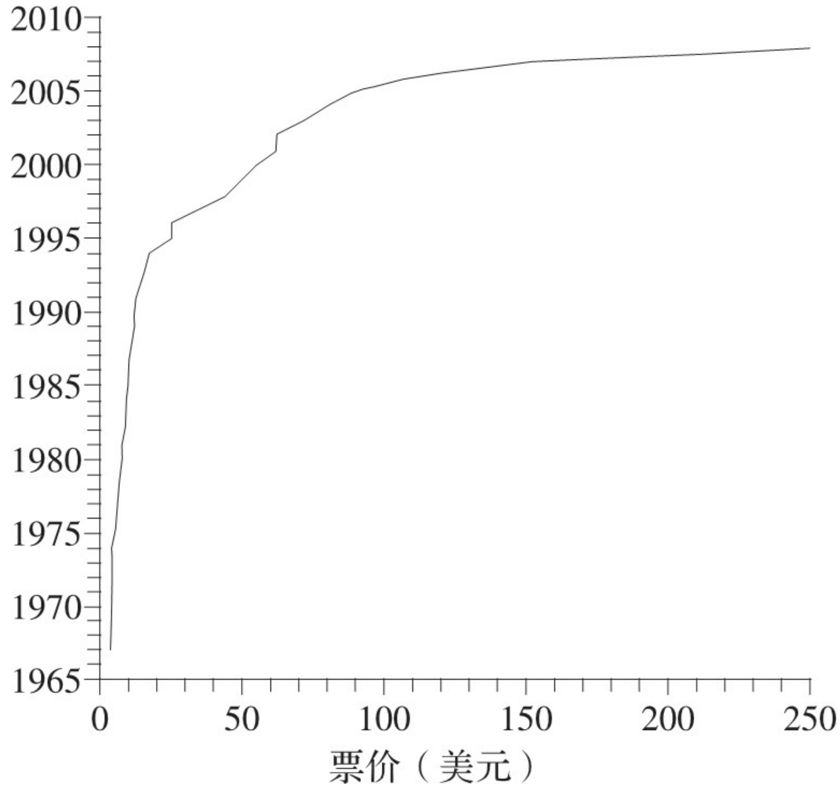


图 5.15　扬基队门票价格增速放缓

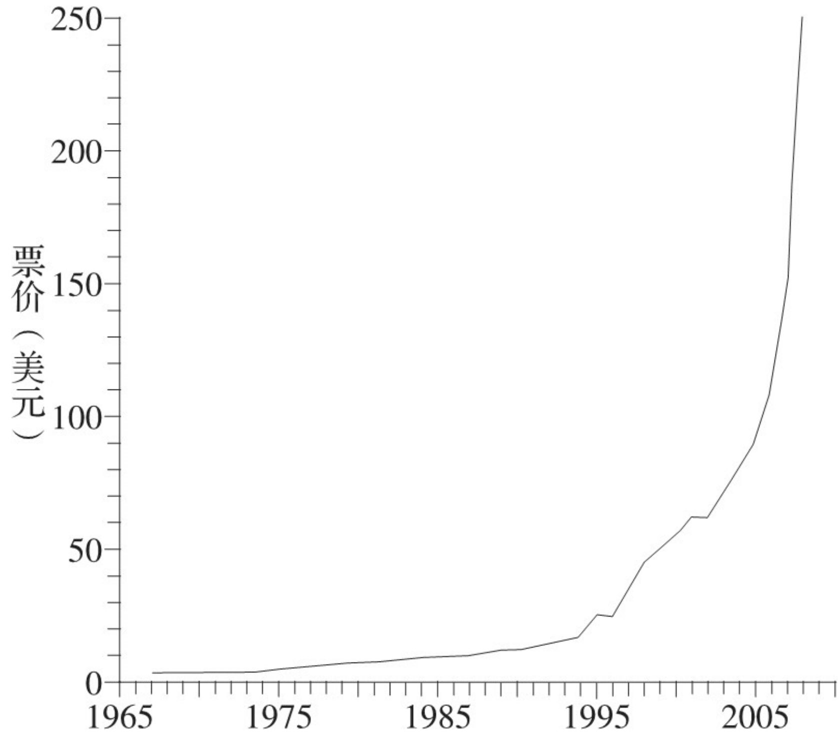


图 5.16　扬基队门票价格加速上涨

图像的制作艺术

《时代》杂志的第一位图像专员是一位艺术学校毕业生，他的目标是“将统计数据表现为某种视觉思想，而不是一堆沉闷的数字。”数字本身并不沉闷。它们可以说明问题，引人注目，甚至令人愉快。当我们认为图像的艺术性比知识性更加重要时，我们就会遇到麻烦。

越大越好？

当吉米·卡特总统（President Jimmy Carter）1979 年任命保罗·沃尔克（Paul Volcker）担任美联储主席时，美国的通货膨胀率已经超过了

13%。在一场针对通货膨胀的全面战争中，美联储将利率提高到了前所未有的水平。在被问及这些货币紧缩政策是否会导致经济衰退时，沃尔克回答道："是的，而且越快越好。"在另一场谈话中，沃尔克表示，"当最后的圆锯停止转动时"，他才会感到满意。这句话的意思是，他希望将利率提到足够高的水平，以堵住借贷源头，使建筑行业停产。

1981 年，住房抵押贷款利率达到了 18%，其他大多数贷款的利率还要更高。随着利率的上涨，家庭和企业减少了借贷以及对汽车、住房和办公楼的购买。失去工作的建筑工人迅速减少了食品、服装和娱乐开支，这给整个经济带来了一种传导效应。面对昂贵的贷款和下降的收入，农场主将拖拉机开进了华盛顿特区中心，封锁了美联储大楼。失业率从 1979 年的 5.8% 上升到了 1982 年的 10% 以上，这也是大萧条以来的最高水平。不过，美联储的焦土政策将通货膨胀率从 1979 年的超过 13% 降低到了 1982 年的 4% 以下。

图 5.17 比较了三位美国总统任期结束时一张百元钞票的购买力。根据定义，在杰拉尔德·福特（Gerald Ford）任期结束时，一张百元钞票的价值为 100 美元。到 1981 年吉米·卡特任期结束时，一张百元钞票只值 67 美元——因为它只能买下四年前用 67 美元买下的东西。到罗纳德·里根任期结束时，一张百元钞票只值 49 美元。



图 5.17　100 美元的价值在不断下降！

这些不断缩水的百元钞票似乎表明，吉米·卡特任职期间钞票价值的下降相对温和，里根任职期间钞票价值的降幅则要大得多——这个结论很奇怪，因为卡特是 20 世纪 70 年代后期严重通货膨胀期间的总统，而里根在 1981 年 1 月才担任总统，当时通货膨胀正在消退。100 美元、67 美元和 49 美元这三个数据显然意味着同里根的任期相比，卡特任期内的美元价值下降得更厉害，但是上面这张图却讲述了一个完全不同的故事。哪个故事才是正确的呢？

图 5.17 的一个问题是，卡特担任了 4 年总统，里根则担任了 8 年总统。虽然这张图考虑到了年份因素，但是这种任期长度的差别并不明显，因为横轴上并没有年份——实际上，图中并没有画出横轴。另一个问题是，由于纵轴的省略，我们无法判断图中是否包含零点。

更糟糕的是，图中没有使用简单的条柱，而是使用了形象的百元钞票。里根任期结束时的物价是福特任期结束时的两倍，百元钞票的价值减少了一半。如果将里根那张钞票的高度减半，同时不改变宽度，图片就会出现失真，就像哈哈镜一样。为避免图片被压扁，里根那张百元钞票的宽度和高度同时被减半——这导致了另一种失真。现

在，里根那张钞票的面积是福特那张钞票的四分之一，因此这张百元钞票的价值看上去减少了四分之三。

这是几个数字比一张图片更加有用的例子之一。卡特任职期间，物价的年增长率是 10.7%，里根任职期间，物价的年增长率是 3.7%——这与用意良好但聪明过度的图 5.17 所传达的错误信息完全不同。

视觉幻象

美洲原住民部落在南加州经营着多家赌场。有一个部落雇用了一位咨询师，让他根据这些赌场与潜在客户的车程对它们的位置进行评估。这是一个复杂的问题，因为你需要考虑客户的居住位置以及他们可以选择的其他赌场。如果一家赌场与一些潜在客户相距 20 公里，那么 60 公里以外另一家赌场对这些客户的吸引力就会受到影响。

咨询师提出了一个专门用于估计“区位价值”的模型。他没有对这个模型做出解释，而是用一张图（图 5.18）展示了模型的结果。这张图没有显示太多信息，是图表垃圾的一个优秀案例。图片、线条、墨水和斑点可以使图像变得更加生动，但它们常常会创造出缺乏吸引力的图表垃圾，使眼睛更加疲劳，使读者更加困惑。

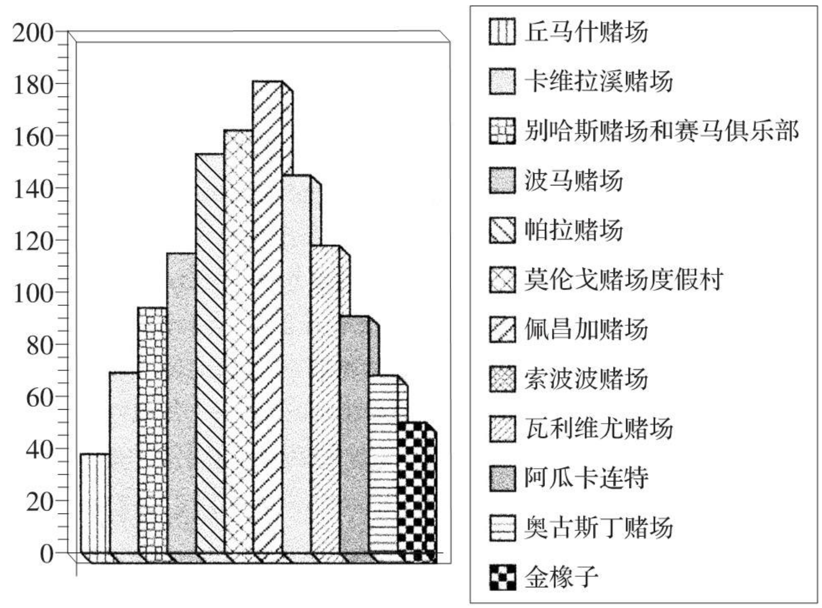


图 5.18　缺乏帮助作用的图像

图像应当显示出无法在表格中明确显示出来的模式，图 5.18 并没有做到这一点。图中的条柱显然表明了这 12 家赌场的区位价值；我之所以使用“显然”一词，是因为纵轴上并没有任何标注。我们并不知道这些数字意味着什么，或者它们是如何计算的。它们是美元吗？是百分比吗？是家庭数量吗？12 个条柱被排列成了钟形曲线，但是作者并没有解释为什么要把丘马什赌场放在最前面，为什么要把金橡子放在最后面。条柱像钟形曲线一样先上升后下降的模式并没有揭示出任何有用的信息。更加明智的做法是把这些条柱按照从高到低的顺序排列。

此外，横轴上并没有任何标注。相反，我们需要在图像和图例之间来回切换，而条柱样式的相似性使这种切换变得乏味而困难。条柱的样式会使人分心，三维外观也没有任何帮助作用。最后，即使我们破译了条柱的含义，我们也不容易将条柱高度与纵轴上的数值相对应。

同表 5.1 这样的简单表格相比，上面这张柱状图并没有表现出更大的价值。有时，简洁就是美。

表 5.1　有时，一张简单的表格胜过一张凌乱的图表

赌场	区位价值
佩昌加赌场	181
莫伦戈赌场度假村	162
帕拉赌场	153
索波波赌场	145
瓦利维尤赌场	118
波马赌场	115
别哈斯赌场和赛马俱乐部	94
阿瓜连卡特	91
卡维拉溪赌场	69
奥古斯丁赌场	68
金橡子	50
丘马什赌场	38

计算机也会生成文本垃圾——将大小、样式和字体不匹配的文字粘贴在一起形成的像勒索信一样的打印文档。文字处理程序中拥有多种样式选项，但这并不意味着我们必须在一份文档中使用所有样式。

我曾收到一份两页纸的时事通讯，上面使用了 32 种不同字体，这还不包括粗体、斜体和不同文字大小所形成的变化。阅读这样的文本是一种痛苦。图表垃圾和文本垃圾都是垃圾，是毫无价值的凌乱，它们无法为人们带来清晰的信息，反而分散了人们的注意力，使人感到困惑。快把垃圾扔掉！

×如何轻松识破一本正经地胡说八道

图像可以揭示某种模式，比如收入随时间的变化以及收入与支出的相互关系。图像也会扭曲数据，误导读者。

当心将数轴上的零点忽略掉的图像。这种忽略可以将图像放大，显示出之前由于分辨率问题而无法发觉的模式。不过，这种做法也会放大数据的波动性，可能产生误导效果。最糟糕的是数轴上没有数字的图像，因为我们无法判断数据的波动性得到了怎样的放大。

当心没有调整人口和物价增长因素的数据。不要被那些将时间放在纵轴上的图像欺骗，尽管我们并不经常见到这样的图像。也不要被间隔不一致的图像欺骗——比如同样的 1 厘米时而表示五年间隔，时而表示十年间隔。

图像不应当仅仅成为一种装饰，为那些缺乏耐心的人带来乐趣。有用的图像可以准确而一致地展示数据，帮助我们理解数据。相反，图表垃圾会分散我们的注意力，使我们感到困惑和烦躁。有些图表垃圾的制作者本意是好的，但是方法不当。有些图表垃圾则是故弄玄虚。

① 传说，当玛丽·安托瓦内特王后听说农民没有面包吃时，她说，让他们吃蛋糕吧。——译者注

第 6 章
美国有多少非裔职业运动员？

伟大的法国数学家皮埃尔-西蒙·拉普拉斯（Pierre-Simon Laplace）曾经评论道，“概率只不过是以计算形式体现出来的常识而已”。我们应当对缺乏思考的计算保持警惕。计算是一件相对容易的工作。更有难度的问题是这种计算是否有道理。

蒙提·霍尔问题

统计难题和悖论是一种具有挑战性的趣味头脑体操。它们也可以说明先思考后计算的价值。蒙提·霍尔问题是一个很好的例子。在电视节目《一锤定音》中，你可以在三扇门之中做出选择。其中，一扇门后面是一项大奖，另外两扇门后面是山羊。在你选择一扇门以后，主持人蒙提·霍尔（Monty Hall）每次都会向你展示你没有选择的一扇门后面的山羊，并且询问你是否想要换一扇门。

大多数人认为，既然还剩下两扇门，那么这两扇门的机会是均等的。不过，请动用一下你的常识。你已经知道你没有选择的两扇门中的一扇门后面是山羊了。“他提醒你这两扇门中的一扇门后面有一只山羊”与“他通过向你展示一只山羊来证明这一点”有区别吗？关于你所选

择的那扇门，你并没有获得任何有用的信息。选择这扇门获胜的可能性仍然是三分之一；因此，选择最后一扇门获胜的概率上升到了三分之二。你应当做出转换。

表 6.1 展示了 300 场游戏的结果，其中第一扇门是你最初的选择。在奖品位于第一扇门后面的 100 场游戏中，主持人一半的时候展示第二扇门，一半的时候展示第三扇门。当奖品在第二扇门或者第三扇门后面时，主持人必须展示另一扇门。

第二扇门被展示了 150 次；第三扇门被展示了 150 次。不管主持人展示的是第二扇门还是第三扇门，奖品位于第一扇门后面的次数都是三分之一。

表 6.1　蒙提·霍尔问题

	第一扇门有奖品	第二扇门有奖品	第三扇门有奖品	总计
第二扇门被打开	50	0	100	150
第三扇门被打开	50	100	0	150
总计	100	100	100	300

对于这个难题，另一种思考方法是假设你在选择第一扇门以后昏了过去。你既没有看到蒙提打开一扇门，也没有听到他询问你是否换一扇门。显然，你获胜的可能性仍然是三分之一。

怀疑者有时会被一百万扇门的极端情况说服。假设你不断进行这项游戏。每一次，蒙提向你展示剩余那些门之中某一扇门后面的山羊。你认为你能够在一半的游戏中获得大奖吗？

凭借“最高智商”入选吉尼斯世界纪录名人堂的玛莉莲·沃斯·莎凡（Marilyn vos Savant）在联合专栏“问问玛莉莲”中讨论了蒙提·霍尔问

题，在全国范围内引发了人们的愤怒抗议。玛莉莲给出了正确答案，随后收到了一万多封信，许多信件来自大学教授，大多数信件认为她说错了。乔治梅森大学的一位数学教授极为愤怒：

> 你搞砸了！让我给你解释一下吧：如果一扇门被证明没有奖品，这种信息会把剩余两个选项的概率变成 1/2。没有任何理由能够使二者之间出现差异。作为一名数学教授，我对公众缺乏数学技能的现象感到深深的忧虑。请帮帮忙，承认你的错误，并在未来的工作中多加注意。

玛莉莲坚持了自己的答案，并且邀请人们在家中进行这项游戏。数万名学生在全国各地的课堂上进行了实验。新墨西哥州洛斯阿拉莫斯国家实验室进行了计算机模拟。渐渐地，舆论发生了转变。玛莉莲是正确的。那位乔治梅森大学的教授表示，“我向她写了另一封信，并在信中告诉她，在收回自己愚蠢的言论之后，我向你低头认错。我发誓，为了赎罪，我将回复所有写信谴责我的人。作为一名教授，我感到极为尴尬。”

关于这段故事，《纽约时报》发表了一篇文章，包括对蒙提·霍尔本人的采访。霍尔曾在数千场《一锤定音》电视节目的结尾主持这项游戏的某个版本。霍尔知道选手最初的选择仍然有三分之一的正确概率，他还知道打开一扇门的做法会使选手认为自己的机会提高到了一半。他将这种做法称为“亨利·詹姆斯处理”或者“旋转的螺丝”。当选手形成“五五开”的思维模式时，霍尔可以为他们提供数千美元转换或不转换的奖励，以便从心理上将他们推向一边或另一边。霍尔还注意到，他不需要按照玛莉莲的规则行动。如果选手最初选择的门后面是山羊，霍尔可以直接打开这扇门，不向选手提供转换机会。霍尔与《纽约时报》撰稿人进行了十场游戏，撰稿人每次都选择了山羊。

一个名叫佛罗里达的女孩儿

另一个悖论是，一个名叫史密斯（Smith）的男人正在和他的女儿散步。史密斯说，他们家还有一个孩子。这个不在身边的孩子是女孩儿的概率是多少？乍一看，这个概率似乎是 1/2。不过，一些专家认为这个回答很天真。这些专家说，正确答案是 1/3。他们还说，如果我们发现和史密斯在一起的女孩儿是他的第一个孩子，那么这个概率就会从 1/3 变成 1/2。这是怎么回事呢？

这个悖论具有许多不同的形式，包括男孩儿、女孩儿、熊、鹰，而且曾经出现在许多不同的场合，包括马丁·加德纳（Martin Gardner）1959 年在《科学美国人》上的一篇专栏文章、约翰·保罗（John Paulos）1988 年的作品《数学盲》以及列纳德·蒙洛迪诺（Leonard Mlodinow）2008 年的作品《醉汉的脚步》。这个问题所依据的传统假设是，在所有二孩家庭中，1/4 家庭有两个男孩儿（BB），1/4 家庭有两个女孩儿（GG），一半家庭拥有一个男孩儿和一个女孩儿（男孩儿先出生的情况记作 BG，女孩儿先出生的情况记作 GB）。这些假设与现实并不完全相符，但是这个题目讨论的是逻辑，而不是数据。

“专家”的观点是，当我们知道史密斯的一个孩子是女性时，只剩下了三种可能性：BG、GB 和 GG。因此，两个女孩儿（GG）的概率是 1/3，一个男孩儿和一个女孩儿（BG 或 GB）的概率是 2/3。这种说法听起来很有道理。不过，在你接受这种观点之前，考虑下面的说法。如果这种逻辑是正确的，那么它也适用于史密斯和男孩儿在一起散步的情况。在这种情况下，我们可以排除 GG 的可能性，认为史密斯拥有一个男孩儿和一个女孩儿的概率是 2/3。

如果这种观点是正确的，那么和史密斯在一起散步的孩子是女孩儿还是男孩儿并不重要！因为在这两种情况下，他拥有一个男孩儿和一个女孩儿的概率都是 2/3。因此，我们不需要知道和史密斯在一起的孩子的性别。这个孩子可以藏在史密斯身后，具有某种模糊的性别，

或者在公园里和其他几十个孩子一起玩耍。不管是哪一种情况，史密斯拥有一个男孩儿和一个女孩儿的概率都是 2/3。这种说法显然是错误的，因为在所有二孩家庭中，只有一半的家庭拥有一个男孩儿和一个女孩儿。常识一定是正确的，专家的推理一定存在缺陷。

回过头来，考虑史密斯与一个孩子在一起散步的各种可能性。表 6.2 显示了在 BB、BG、GB 和 GG 之间均匀分配的 400 个家庭。在史密斯有两个男孩儿的 100 种情况中（BB），他总是和一个男孩儿散步。在史密斯有两个女孩儿的 100 种情况中（GG），他总是和一个女孩儿散步。在他拥有一儿一女的情况中（BG 或 GB），一个合理的假设是，他与男孩儿或女孩儿散步的概率相等。

表 6.2　史密斯的另一个孩子是男孩儿或女孩儿的概率相等

	BB	**BG**	**GB**	**GG**	总计
和女孩儿散步	0	50	50	100	200
和男孩儿散步	100	50	50	0	200
总计	100	100	100	100	400

现在观察第一行，即史密斯和女孩儿散步的 200 种情况。在 100 种情况中（GG），不在场的孩子是女孩儿，在另外 100 种情况中（BG 或 GB），不在场的孩子是男孩儿。在第二行里（史密斯和男孩儿散步的 200 种情况），在 100 种情况中（BB），不在场的孩子是男孩儿，在另外 100 种情况中（BG 或 GB），不在场的孩子是女孩儿。不管和史密斯散步的孩子是女孩儿还是男孩儿，他的另一个孩子是男孩儿或者女孩儿的概率都是相等的。专家错了，常识是正确的。

专家们根据错误的逻辑继续说道，如果我们知道和史密斯散步的女孩儿比另一个孩子年长，那么另一个孩子是女性的概率就会从 1/3 提升到 1/2。专家的理由是，在我们知道陪伴史密斯的女儿是他的第一个孩子以后，我们可以排除 BB 和 BG 的情况，只留下 GB 和 GG 的情况。因此，两个女儿的概率从 1/2 提升到了 1/2。不过，让我们用常识考虑一下这种逻辑的推论。如果知道女儿年长可以将两个女儿的概率从 1/3 提升到 1/2，那么知道女儿年幼也可以将两个女儿的概率从 1/3 提升到 1/2。不过，这个女孩儿一定不是年长就是年幼。根据专家的说法，在这两种情况下，两个女儿的概率都会从 1/3 提升到 1/2。因此，即使我们不知道这个女孩儿年长还是年幼，我们也知道两个女儿的概率是 1/2！常识仍然是正确的，专家仍然是错误的。

最近，蒙洛迪诺改进了二孩悖论，认为如果这个女孩儿宣布自己拥有一个独特的名字，比如佛罗里达，那么这种做法也会将两个女孩儿的概率从 1/3 提升到 1/2。他的理由仍然与直觉相反，而且最终被证明是不正确的。如果蒙洛迪诺的说法是正确的，它将适用于每一个名字，因为每个名字都是独特的。（如果教名不够独特，可以使用教名和中间名，或者使用教名、中间名和出生日期。）如果这种说法适用于每一个名字，那么这个名字是什么并不重要，我们是否知道这个名字也不重要。这里没有悖论，只有受到歪曲的逻辑。

2010 年，在两年一度纪念马丁·加德纳的“加德纳集会”上，加里·福希（Gary Foshee）提出了这个问题的另一个版本。他走上讲台，说道：“我有两个孩子。一个是男孩儿，出生在星期二。我有两个男孩儿的概率是多少？”停了一会儿，福希继续说道：“你能想到的第一件事情是，‘这和星期二有什么关系？’实际上，二者之间存在密切的关系。”然后，福希走下了讲台。他的发言在会场和互联网上引发了一场热烈的讨论。

我的回答是，这和星期二的确没有任何关系。如果星期二能够改变这个概率，那么星期三、星期四或者一周里的其他任何一天也能以同样的方式改变这个概率。不过，这个孩子一定会出生在一周里的某

一天。因此，如果福希的说法是正确的，我们可以在不知道这一天是星期几的情况下改变这个概率。福希是错误的。这一天是星期几并不重要。

条件概率的混淆

我有一个缺乏耐心的亲戚——让我们叫他鲍勃（Bob）吧。一次，鲍勃走进一家大卖场，准备购买一部智能手机。手机的款式实在是太多了！他请一位店员帮忙。面对店员说出的令人难以理解的话语，鲍勃的沮丧感不断加深。蓝牙、地理标记、兆像素、HDML、IOS、LCD、RAM、ROM，等等。最终，鲍勃留下一句“你嗑药了吧”，然后转身离去。

在开车回家的路上，鲍勃想到了一个好主意。企业应当通过药检筛选求职者，监督员工。奥运会有药检，自行车运动有药检，赛狗运动也有药检，为什么不能把药检运用到某种重要的事情上，比如服务顾客呢?

鲍勃做了一些调查，发现针对吸食大麻的简单尿检拥有 95% 的准确率。这当然已经足够好了！既然 95% 的准确率在统计检验中是一个足够好的准确率，那么它当然也可以用于排除瘾君子。

不过，如果将鲍勃的想法付诸实践，美国的失业率可能会急剧上升，而且这并不是因为我们是一个毒品之国。鲍勃犯了一个很常见的错误，他混淆了两种条件概率。

几年前，一位黑人大学教授询问一群黑人退伍军人：美国有多少非裔职业运动员？退伍军人给出的猜测在 5 万人到 50 万人之间。正确答案是 1200 人。美国的黑人律师是这个数字的 12 倍，黑人医生是这个数字的 15 倍。没有一个退伍军人相信他的话，但他的说法是正确的。

这一现象的深层次问题是，人们将不同的条件概率混淆在了一起。我们看到黑人在职业运动员中占有很大的比例，因此下意识地认为很大一部分黑人都是职业运动员。如果我们将话题从种族转变成性别，这个错误就会更加明显。美职篮的所有球员都是男性，但是男性之中在美职篮打球的群体只占很小的一个比例。

当年，许多（也许是大多数）非裔美国孩子希望成为下一个迈克尔·乔丹（Michael Jordan）或者魔术师约翰逊（Magic Johnson）。今天，他们希望成为勒布朗·詹姆斯（LeBron James）或者凯文·杜兰特（Kevin Durant）。遗憾的是，他们几乎一定无法成功。拥有梦想和抱负是好的，但是认清现实更加重要。学术和体育运动都很重要，学者兼运动员应当受到尊重。不过，同运动员相比，学者更有可能得到一份好工作。

2012 年，俄亥俄州立大学三线四分卫发布了一条推文："我们是来打橄榄球的，为什么我们还要上课，我们不是来打学校的，上课是没有意义的。"好吧，几年以后，我们再来看看情况如何。

假阳性问题

毒品检测领域也存在同样的混淆问题。95% 的大麻检测准确率意味着在使用大麻的人之中，95% 的人会检测出阳性。那么，问题来了：在检测出阳性的人之中，大麻使用者的比例是多少？

如果对员工进行毒品检测，会产生两类错误。如果检测错误地发现毒品痕迹，这种错误叫做假阳性；如果检测没能发现毒品痕迹，这种错误叫做假阴性。为了说明假阳性可能产生多大的问题，考虑一项针对 1 万名员工的检测，其中 500 名员工（5%）使用了大麻，9500 名员工（95%）没有使用大麻。进一步说，假设检测的准确率是

95%：95% 的大麻使用者会得到阳性检测结果，95% 不使用大麻的人会得到阴性检测结果。

表 6.3 显示，在使用大麻的 500 人中，475 人（95%）得到了阳性检测结果，25 人（5%）没有得到阳性结果。在不使用大麻的 9500 人中，475 人（5%）得到了阳性检测结果，9025 人（95%）没有得到阳性结果。到目前为止，情况还算不错。

表 6.3 假阳性问题

	检测呈阳性	检测呈阴性	总计
大麻使用者	475	25	500
非大麻使用者	475	9025	9500
总计	950	9050	10000

不过，表 6.3 还显示，在 950 个阳性检测结果中，475 个结果是假阳性。高达 50% 的阳性员工没有使用大麻。这就是我们需要谨慎对待条件概率的原因。虽然 95% 的大麻使用者会检测出阳性结果，但是只有 50% 的阳性结果来自大麻使用者。

在 1997 年的"钱德勒诉米勒案"中，美国最高法院以 8 比 1 的表决结果裁定佐治亚州要求某些州政府岗位申请者接受毒品检测的法律，违反了第四修正案反对不合理调查的规定。法院认为，在具体个体没有受到犯罪怀疑但某种特殊需要优先于隐私权的一些要求"严密保护"的情况下，调查是被允许的——比如航班飞行员的毒品检测。不过，法院认为，州级当选官员使用毒品的可能性不会危害公共安全。（你可以在这里吐槽了。）

罕见病问题

假阳性问题也存在于针对疾病的医学检测之中。和大麻检测类似，即使针对疾病的检测拥有很高的准确率，得到阳性检测结果的许多人（甚至大多数人）仍然有可能不是这种疾病的患者。

下面的例子很能说明问题。一百位医生被问到了这样一个虚拟问题：

> 在一次常规检查中，你在一位女性患者的乳房上发现了一个肿块。根据你的经验，在 100 个这样的肿块里，只有 1 个肿块是恶性的。不过，为了安全起见，你要求患者接受乳房 X 光检查。如果肿块是恶性的，那么 X 光检查将其诊断为恶性的概率是 0.80；如果肿块是良性的，那么 X 光检查将其诊断为良性的概率是 0.90。在这个例子中，X 光检查认为肿块是恶性的。根据这个 X 光检查结果，你认为这个肿块是恶性肿块的概率是多少?

在 100 位接受调查的医生中，95 位医生给出了 0.75 左右的概率。不过，正确的概率是这个数字的十分之一：0.075！

表 6.4 显示了 1000 名患者的情况。在 10 个病例中（1000 人的 1%），肿块是恶性的。在 8 个恶性病例中（80%），检测给出了正确的阳性结果。在 990 个良性病例中，检测给出正确阴性结果的病例为 891 个（90%）。

表 6.4　乳房 X 光检测的假阳性

	检测呈阳性	检测呈阴性	总计

恶性	8	2	10
良性	99	891	990
总计	107	893	1000

观察第一行数据。在10名拥有恶性肿瘤的病人中，检测给出阳性结果的次数是80%：8/10=0.80。不过，观察第一列数据。在得到阳性检测结果的107名患者中，只有7.5%的患者拥有恶性肿瘤：8/107=0.075。尽管10个恶性肿瘤中的80%得到了正确诊断，这些阳性检测结果仍然远远少于假阳性结果——990个良性肿瘤中被错误诊断的10%。正像这里展示的那样，有时大数中的小比例大于小数中的大比例。

条件概率很容易得到错误解读，而这些医生显然犯了这个错误。开展这项调查的研究人员指出，

> 犯错误的医师常常表示，他们认为病人在得到阳性X光结果的情况下患癌的概率……与癌症患者得到阳性X光结果的概率大致相等……后一种概率是临床研究计划中测量到的概率，是医生熟悉的概率；前一种概率则是制定临床决策时需要使用的概率。看起来，许多甚至大多数医师并没有分清这两种概率。

显然，大多数医生混淆了下列条件性说法。如果肿块是恶性的，检测结果呈阳性的概率是多少？（答案：80%。）如果检测结果呈阳性，肿块是恶性肿块的概率是多少？（答案：7.5%。）这些医生对医学数据的错误解读可能导致灾难性后果。

达特茅斯三文鱼研究

当检测数量很多时，假阳性是不可避免的。例如，假设一个看上去很健康的女性接受体检，包括多项可能指示健康问题的风险因素独立检测（比如胆固醇和高血压）。在每项检测中，如果读数超出 95% 健康女性的读数范围，结果就会被标为“异常”。在每一项检测中，假阳性的概率是 5%。对于十项检测，假阳性的概率是 40%。对于一百项检测，假阳性结果的概率为 99% 以上。随着检测数量的增长，假阳性的概率将接近 100%。

下面的例子涉及一项包含大量检测的标准神经科学实验。实验人员将志愿者放进磁共振成像仪，向他展示各种图像，并且提出关于这些图像的问题。这种实验不是用植入大脑的电极跟踪大脑活动，而是用功能磁共振成像（fMRI）测量含氧和脱氧血液流过大脑时产生的磁干扰。检测过后，研究人员观察超过 13 万立体像素（三维数据），以查看大脑的哪些部位受到了图像和问题的刺激。

功能磁共振成像的测量包含许多噪声，包括来自环境以及来自大脑不同部位脂肪组织密度差异的各种磁信号。有时，立体像素会忽略大脑活动（假阴性）。有时，立体像素会错误地指示大脑活动（假阳性）。

一个名叫克雷格·贝内特（Craig Bennett）的学生，在达特茅斯实验室以一种独特的方式开展了这项实验。他用磁共振成像仪研究一条三文鱼在依次看到 15 张照片时的大脑活动。下面是实验报告的一部分：

对象。参与功能磁共振成像研究的一条成熟的大西洋三文鱼（大西洋鲑）。这条三文鱼长约 45 厘米，重约 1.7 公斤，在扫描时处于无生命状态。

任务。对相关三文鱼执行的任务包括完成一个需要动用大脑的开放式任务。实验向三文鱼展示了一系列照片，照片上描述了社会情境中具有特定情绪效价的人类个体。实验要求三文鱼确定照片中的个体正在经历的情绪。

设计。刺激以区组设计的形式呈现，每张照片展示 10 秒，然后休息 12 秒。共有 15 张照片得到展示。总扫描时间为 5.5 分钟。

分析。实验用一般线性模型（GLM）的普通最小二乘估计对三文鱼立体像素数据进行了处理，用方脉冲函数与标准血液动力反应的卷积模拟血液动力反应的预测指标，并且添加了一个 128 秒的时域高通滤波器，以校正低频漂移。实验没有使用自相关校正。

分析部分听上去很专业，不是吗？不过，你是否在对象部分注意到这条三文鱼“在扫描时处于无生命状态”？没错，克雷格在当地市场买了一条死三文鱼，放进磁共振成像仪，向它展示照片，然后向它提出问题。不过，由于立体像素数量众多，因此出现了一些假阳性，可以被解释成三文鱼对照片和问题的反应。只是这条三文鱼已经没有了生命。

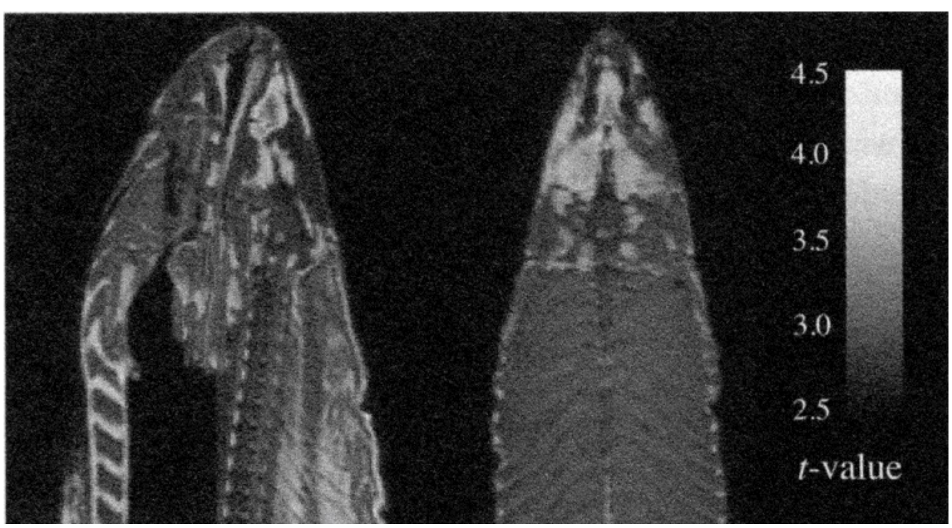


图 6.1　三文鱼的功能磁共振成像检测结果

根据这项实验，贝内特和他的教授艾比盖尔·贝尔德（Abigail Baird）提出了一个有力的观点：功能磁共振成像研究需要考虑到假阳性问题。多达 40% 的已发表论文并没有做到这一点。

这项死三文鱼研究获得了超过大多数功能磁共振成像研究的曝光度，甚至获得了搞笑诺贝尔奖——哈佛大学每年都会举办一个非常欢乐的颁奖仪式，以奖励“首先使人发笑，然后使人思考的成就”。

×如何轻松识破一本正经地胡说八道

蒙提·霍尔问题是一个绝妙的悖论，因为我们的直觉是错误的，而且可以通过运用常识得到改进。二孩悖论则是另一种类型的精彩悖论。在这个悖论中，我们的直觉是正确的，而且可以通过使用常识得到证实。

不要仅仅进行计算。运用常识思考你所回答的问题是否正确，假设是否合理，结果是否可信。如果一种统计观点不合理，应对其进行仔细思考；你可能会发现，这种观点是在胡说八道。

假阳性问题与条件概率的混淆有关。在某些情况下（比如存在某种疾病），一项检测很有可能显示阳性结果，但阳性检测结果并不能认定疾病的存在。它可能是假阳性。对于罕见疾病（比如恶性肿瘤）或者存在大量读数的情形（比如死三文鱼磁共振成像），假阳性现象更为常见。

第 7 章
辛普森悖论

霍乱是一种以腹泻和呕吐为标志的肠道疾病，常常会导致死亡。霍乱最初仅仅存在于印度次大陆。不过，19 世纪陆地和海洋贸易路线的发展将这种疾病传播到了世界各地，导致数千万人死亡。

1832 年，霍乱袭击伦敦，导致 6500 人死亡。当时的医疗机构认为，霍乱和其他疾病一样，是由呼吸“瘴气”（有毒气体）引起的。腐烂的垃圾、街道上的人畜粪便以及受到污染的泰晤士河散发出来的臭味使伦敦臭气熏天。在潮湿多雾的夜晚，空气尤其令人讨厌。许多人惧怕“夜晚的空气”，他们躲在门窗紧闭的室内。如果不得不外出，他们也会挡住面部。在缺乏像样的卫生服务、味道难闻的贫困街区，霍乱更加常见，这一现象支持了瘴气理论。

瘴气理论无法得到明确证明，因为这个问题存在其他一些需要考虑的混杂因素。例如，整体来看，贫困街区居民的年龄高于其他街区。一些人吃着不同的食物，从事着不同的行业，缺少甚至没有暖气。真正发挥作用的是其中的一个因素还是所有因素呢?

有时，科学家可以在受控条件下进行实验，使混杂因素保持恒定，以分离出一个因素的影响。如果其他所有相关因素维持恒定，那么观测到的结果显然是由变化的因素导致的。不过，这种理想的对照实验常常是不现实或不道德的，因此研究人员必须根据观察到的现象得出结论，而无法对其进行控制。霍乱问题即是如此。医生当然无法

强制性地随机选择一部分人呼吸有毒气体，同时随机选择另一部分人在空气清洁的环境下生活和工作。不过，当时出现了一个自然形成的实验，彻底揭示了这种疾病的原因。

几个世纪以来，伦敦和其他大城市的住宅用地下粪坑收集人体废物。专业人员会定期将粪坑清空，并用运货马车将废物运走。这些废物被称为“夜间土壤”，因为夜间的城市街道行人稀少，专业人员会选择在这个时候将其运走。通常，这些废物被运往附近的农场，以充当肥料。

1848 年，一部法律要求伦敦市民停止使用粪坑，将他们的住宅与遍布整个伦敦的新建污水管道相连。这部法律减少了粪坑和“夜间土壤”运输所导致的恶臭和危险，但它也制造了另一个问题。污水管道将未经处理的污水排入泰晤士河，而这条河流是许多伦敦市民的直接或间接饮用水来源。

1849 年，36 岁的医生约翰·斯诺（John Snow）发表了一篇论文《论霍乱的传播模式》，认为霍乱不是由呼吸污浊空气导致的，而是由“患病人体排出的、通过被污水污染的饮用水传播的一种毒物”导致的。我们并不知道斯诺是如何提出这种理论的。也许，他认为呼吸糟糕的空气应当影响肺部，但霍乱影响的却是肠道，这意味着它与人们的饮食有关。斯诺还发现，在 1848 年禁止粪坑的法律颁布后不久，伦敦就发生了 1848～1849 年的霍乱大流行。当然，这种推理是一种被称为“前后即因果”的逻辑谬论。一个事件紧随另一个事件发生并不意味着后面的事件是由前面的事件导致的。

弗洛伦斯·南丁格尔（Florence Nightingale）和其他顶级公共卫生权威认为斯诺的想法过于天真，缺乏依据。瘴气理论已经根深蒂固，很难被一个年轻医生的奇异猜测所推翻。斯诺无法固定包括瘴气在内的所有潜在混杂因素，强迫一些人饮用被污染的水，强迫另一些人饮用干净的水，以检验他的理论。不过，在 1854 年，当另一场流行霍乱袭击伦敦时，斯诺想出了检验这种理论的两种不同的方法。

多年来，萨瑟克和沃克斯豪尔水务公司与兰贝斯公司一直在通过不同的管道将泰晤士河同一受污染区域的水源输送到伦敦的相同街区。在 1848 ~ 1849 年霍乱流行期间，两家水务公司的客户拥有相同的死亡率。在 1848 年禁止粪坑的法律颁布后，英国微生物学家亚瑟·哈索尔（Authur Hassall）对伦敦水源与污水系统之间的关系进行了详细而充分的研究。他所得到的令人震惊的结果，1850 年发表在英国著名医学期刊《柳叶刀》以及一部名为《对伦敦和郊区居民水源的一项显微镜研究》的书籍中，他写道:

> 我已经证明，泰晤士河水中始终可以检测到与污水相连的各种动植物物质，包括一些粪便；而且，同样的物质存在于一些公司向公众提供的水源之中；这条证据链是完整而具有决定性的；因此，我们反复发现，肉类的肌纤维以及被食用的蔬菜组织的更多无法破坏的部位从抽水马桶进入污水管道，从污水管道进入泰晤士河，从泰晤士河进入水务公司的蓄水池，从蓄水池重新回到公众家中。
>
> 因此，毫无疑问，根据目前的伦敦供水系统，这座城市的一部分居民不得不以某种形式消费自己的排泄物，并为这种特权支付账单。

在哈索尔详细而令人担忧的结论的影响下，1852 年的《大都会水法》规定，从 1855 年 8 月 31 日起，伦敦水务公司不得从泰晤士河严重污染的区域采水。兰贝斯公司已经获得了上游 35 公里处的土地，并在 1852 年做出了改变，开始从没有受到伦敦污水影响的泰晤士河区域采水。萨瑟克和沃克斯豪尔公司直到 1855 年才开始搬迁。

斯诺意识到，这是对他的理论进行检验的绝佳时机。相邻住宅的居民恰巧由不同的水务公司提供服务，这一事实自然而然地控制住了各种潜在混杂因素。斯诺写道:

同时，这项实验也具有最大的规模。不同性别、各个年龄段和职业、各种阶层和地位、从上流人士到穷人的 30 多万市民毫无选择地被划分成了两个群体，而且在大多数情况下，他们对此并不知情；一个群体的供水系统包含伦敦污水，其中含有可能来自霍乱病人的任何物质，另一个群体的供水系统中基本没有这种杂质。

斯诺考察了 1854 年霍乱流行前 7 个星期的所有病人死亡记录，并且确定了由这两家水务公司提供水源的家庭。他发现，萨瑟克和沃克斯豪尔公司每一万户家庭的死亡数字是另一家公司的将近 9 倍:

表 7.1（伦敦各地区霍乱致死人数对比）

	家庭数量	霍乱死亡数量	每一万户家庭的死亡数量
萨瑟克和沃克斯豪尔公司	40046	1263	315
兰贝斯公司	26107	98	37
伦敦其他地区	256423	1422	59

这些数据令人信服地证明了饮用被污染的水与霍乱发病之间的关系。

斯诺还发现了支持这种理论的其他有力证据。1854 年的流行霍乱对索霍区的影响尤其强烈，在 10 天之内导致 500 多人丧生。斯诺就住在索霍区附近。3/4 的居民逃离了这个地区，显然是想躲开这里的瘴气。不过，斯诺并没有离开，他想对自己的理论进行调查。

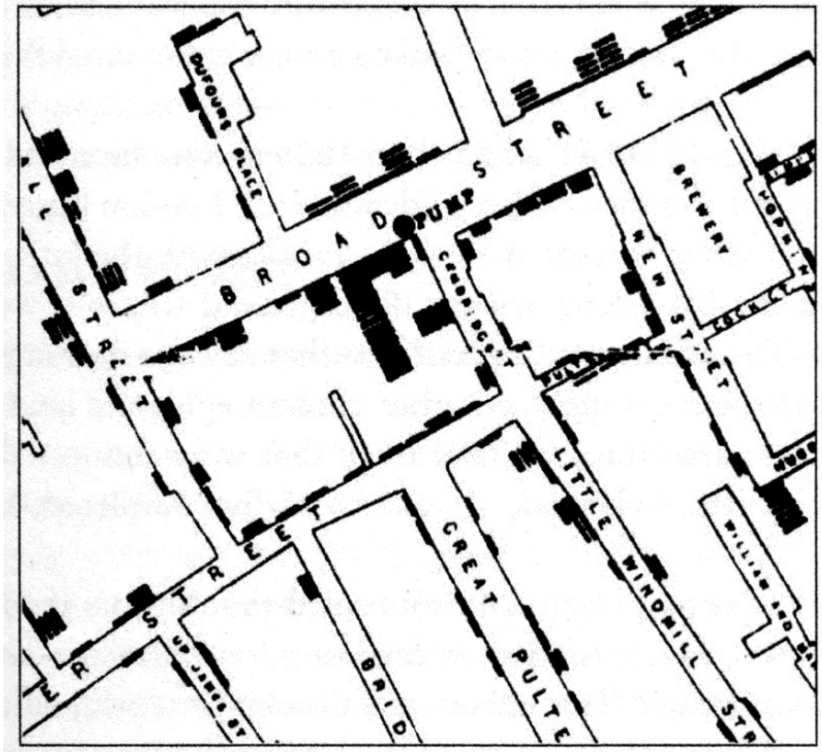


图 7.1　布罗德大街水泵附近的霍乱死者，用住址前的横线表示。

当时，整个伦敦市有几十口公共水井，人们可以在这里打水喝，或者将水带回家。索霍区还没有与伦敦污水系统相连接，斯诺怀疑索霍的粪坑正在污染从公共水井打上来的水。斯诺画了一张图，显示了13个公共水泵的位置以及578个霍乱受害者的住所。他很快发现，许多受害者生活在布罗德大街的一个水泵附近。这个水泵位于布罗德大街和剑桥大街的十字路口。图中的叠加线显示了霍乱受害者的住址，这些受害者显然集中于水泵附近，很可能是因为他们喝了这里的井水。

斯诺亲自走访了生活在其他公共水泵附近的霍乱受害者的家庭，发现他们常常在工作、购物或上学路上经过布罗德大街的水泵时饮用这里的井水。斯诺还发现，一些家庭之所以使用布罗德大街的水泵，是因为这里的水味道比较好。

另外，许多生活在布罗德大街水泵附近但不喝井水的人并没有受到霍乱的影响，这给瘴气理论以沉重一击。这个街区的人们呼吸着相同的空气，但是只有用布罗德大街的水泵喝水的人才会死于霍乱。

圣詹姆斯教区负责管理布罗德大街水井的监护委员会急于采取措施结束这场流行霍乱。斯诺说服他们取走了水泵把手，以阻止人们饮用这里的井水。霍乱很快停止了。不过，在霍乱停止以后，人们重新安上了水泵把手。瘴气理论过于根深蒂固，很难被一位年轻的医生推翻。

后来，人们发现了布罗德大街水泵传播霍乱的原因。最初，这里曾经是一座住宅，一个患上霍乱的婴儿的尿布曾被丢进住宅的粪坑里。当房屋被烧毁、街道被拓宽时，房子里的粪坑遭到了遗弃。人们在距离粪坑 0.9 米的地方挖了这口水井。后来，粪坑发生了渗漏，污染了布罗德大街的水井。

1858 年，斯诺患上了中风，不幸病故。他没能看到自己的理论被人接受，没能看到流行霍乱的终结。1859 年，伦敦开始建设一个现代污水系统。具有讽刺意义的是，这项措施的目的是将污水管线布置在地下，以免其散发气味，从而减少瘴气。这种做法无意中带来了一个好处，那就是污水不再污染人们的供水系统。1866 年，另一场流行霍乱袭击伦敦，但是与新的污水系统相连接的城镇区域并没有受到影响。当这个系统彻底完工时，伦敦终于安全了。

现代科学家最终证明了斯诺的理论。霍乱是通过摄取被粪便物质污染的饮食传播的。所有发达国家的城市都建设了有效的污水处理系统和清洁供水系统，消灭了流行霍乱。今天，约翰·斯诺以其在研究重要公共卫生问题时对数据的独创性使用得到了认可和纪念。他被视为研究疾病模式、原因和影响的流行病学之父。

混杂因素常常出现在使用观测性数据的研究中，因为人们无法通过现实的方法使这些因素维持恒定。不过，大自然有时会为研究人员提供便利条件。在霍乱研究中，斯诺非常幸运地研究了伦敦市相邻住宅由不同水务公司服务的区域，从而减少了社会经济因素的混杂效

应。斯诺指出，生活在相邻住宅的家庭可能来自相同的社会经济阶层。如果斯诺需要对不同水务公司服务的不同街区或城市进行比较，这种条件就无法成立了。

有时，为了处理混杂因素，可以对观测性数据进行细分。在研究吸烟对健康的影响时，性别可能是一个混杂因素。一项得到良好设计的研究可能会分别分析男性和女性的数据，以控制性别因素。在研究收入对生育的影响时，年龄和宗教信仰可能成为混杂因素。一项得到良好设计的研究可能会把数据分成具有相同年龄和相似宗教信仰的不同小组。

最重要的一点是，我们应当永远牢记，一项研究的结论有可能受到混杂因素的干扰。

某人的悖论

2010 年，一家互联网公司收集了两种不同网页布局的数据。在“一次点击”形式中，广告出现在网页的第一个页面上。在“二次点击”形式中，第一个页面上显示的是关键词；如果用户点击关键词，与这个关键词有关的定向广告就会显示出来。在这两种形式中，如果用户点击广告，公司就会获得收入。二次点击形式要求用户付出更多精力，但定向广告可能更加有效。因此，当用户选择“二次点击”广告时，公司可以获得更多收入。

分析人员向公司总裁展示了表 7.2 所示的数据。（收入和用户数的单位均为百万。）二次点击形式具有较高的 RPM（平均一千名用户的收入），为 12.14 美元，一次点击则只有 11.60 美元。下一步的行动似乎很明显：如果他们在所有网页上使用二次点击形式，收入就会增长。

表 7.2　收入、用户与千人收入（RPM）

一次点击			二次点击		
收入	用户数	RPM	收入	用户数	RPM
$2.9	250	$11.60	$1.7	140	$12.14

这个结论可能是一个昂贵的错误。这些数据属于观测性数据，可能存在自选择偏差，即访问“一次点击”网站的用户可能与访问“二次点击”网站的用户存在系统性差异。要想进行有效的比较，公司可以进行一项对照实验。每个网站可以将用户随机导向两种形式中的一种。这样一来，公司就可以对每一种形式的效果进行有效的比较。

还有一个不太明显的问题。一个讨厌的统计学家分发了一张表（表 7.3），这张表对美国本土和国外的用户进行了划分。在两种用户中，一次点击形式都具有更高的 RPM。不出所料，人们感到极为震惊。总裁举起双手，询问为什么一次点击形式在美国本土和国外全部占优，但在总体上却处于劣势。一些员工提议对这些数据进行检查。其他一些人露出茫然的表情，希望某人能够站出来解决这个悖论。

然后，大家明白过来了。

表 7.3：区分美国用户与国际用户

	一次点击			二次点击		
	收入	用户数	RPM	收入	用户数	RPM
美国	$1.8	70	$25.71	$1.2	50	$24.00
国际	$1.1	180	$6.11	$0.5	90	$5.56

总计	$2.9	250	$11.60	$1.7	140	$12.14

这是辛普森悖论的一个例子。虽然爱德华·辛普森（Edward Simpson）在 1951 年的一篇论文中描述了这个悖论，但它实际上是由另外两位统计学家在 50 年前发现的，这使辛普森悖论成了斯蒂格勒定律的一个例子。斯蒂格勒定律的内容是："没有一项科学发现是以其最初发现者的名字命名的。"（斯蒂格勒本人就是一个例子，他指出，罗伯特·K.默顿才是斯蒂格勒定律的发现者。）辛普森悖论指的是当聚合数据被分解时其中的模式发生逆转的现象。在上面的例子中，对于聚合数据来说，二次点击占据优势。当数据被分解成美国和国际两部分时，一次点击占据优势。

要想理解这种逆转，首先考虑二次点击网站拥有较高 RPM 的聚合数据。用户类型（美国或国际）是一个混杂因素，因为 RPM 不仅与点击形式有关，也与用户类型有关。同国际用户相比，美国用户拥有更高的 RPM，而且恰巧更喜欢访问二次点击网页，这推高了二次点击的总体 RPM。如果我们将这种混杂因素考虑在内，将美国用户和国际用户的数据分开，我们就会发现，在两种类型中，一次点击拥有更高的 RPM。

如果所有网页使用一次点击形式，公司的总体收入才有可能提高，这与他们最初的印象相反。幸运的是，公司的政策专家非常聪明，认识到了辛普森悖论，而且进行了上面描述的对照实验。他们的结论是，一次点击形式在大多数网站上表现得更好，但是一些网站更适合二次点击形式。

要想注意到可能的辛普森悖论，关键是考虑是否存在被忽略的混杂因素。下面是另一个例子。20 世纪 70 年代，有人指控加州大学伯克利分校研究生院歧视女性申请人。作为证据，他们提供了表 7.4 中的数据，指出男性申请人的录取率为 44%，而女性申请人的录取率只有 35%。

表 7.4：男性更有可能被录取

	申请人	录取率
男性	8442	44%
女性	4321	35%

法院启动了一项调查，以确定哪些系的问题最为严重。不过，在考察了该学院 85 个系的录取率以后，人们几乎没有发现女性受到歧视的证据。相反，他们觉得一些系对于女性的录取率反而高于男性。

表 7.5 显示了最大的六个系的录取率。第一个系拥有最高的录取率，第二个系拥有第二高的录取率，依此类推。总体而言，男性申请人的录取率是 45%，女性申请人的录取率则只有 30%——这似乎是歧视女性的明显证据。不过，在考察每个系时，只有两个系（第三个系和第五个系）男性的录取率高于女性，而且这种差异很小，不具有统计显著性。唯一具有统计显著性的录取率差异出现在第一个系。在这里，女性的录取率明显高于男性（82% 和 62%）。

表 7.5：最大的六个系的录取率

	总计		男性		女性	
系	申请人	录取率	申请人	录取率	申请人	录取率
1	933	64%	825	62%	108	82%
2	585	63%	560	63%	25	68%
3	918	35%	325	37%	593	34%
4	792	34%	417	33%	375	35%

5	584	25%	191	28%	393	24%
6	714	6%	373	6%	341	7%
总计	4526	39%	2691	45%	1835	30%

这仍然是辛普森悖论。当数据被分解时，聚合数据中的模式遭到了逆转。这里的混杂因素是，一些系的录取率远高于其他系。第一个系拥有 64% 的总录取率；第六个系拥有 6% 的总录取率。现在，你应该注意到，男性更喜欢申请第一个系，而不是第六个系，女性则恰恰相反。

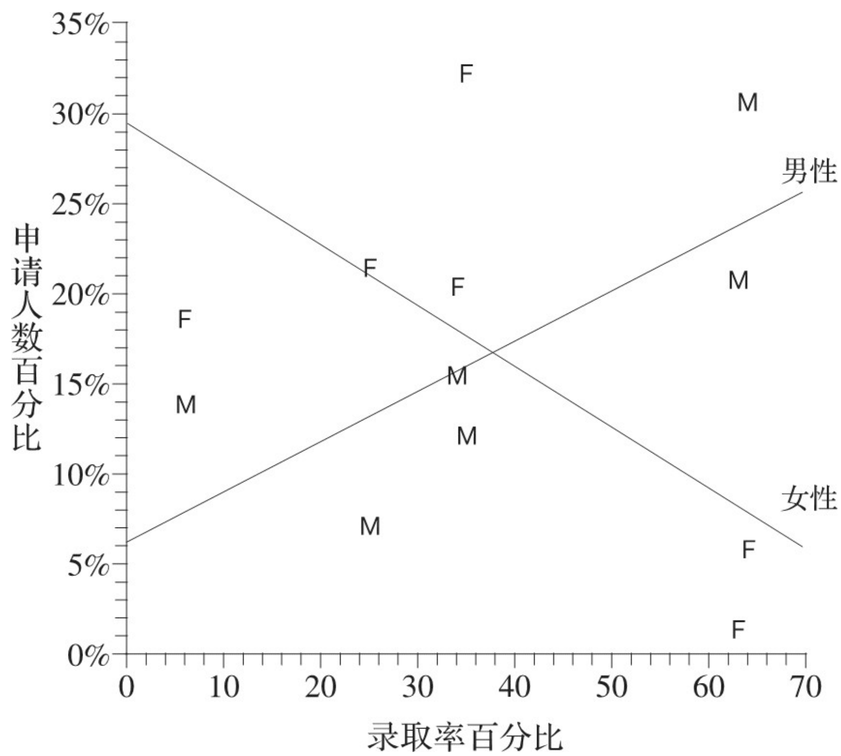


图 7.2　女性更喜欢申请录取率较低的系

图 7.2 以图形的方式证明了表 7.3 所显示的结论。横轴是系录取率。纵轴是每个性别申请这个系的人数比例。字母 F 代表女性申请人，M 代表男性申请人。例如，最左边的两个点代表第六个系，它的录取率是 6%。14% 的男性申请人和 19% 的女性申请人申请了第六个系。两条拟合线表明，总体而言，女性倾向于申请录取率较低的系，男性申请人则恰恰相反。男性的总体录取率较高，因为他们以不成比例的人数申请了最容易进入的系。在考察了所有 85 个系并且考虑到这种混杂因素以后，人们撰写了一份详细的研究报告，指出该校“对女性存在微小但具有统计显著性的偏向”。

发现混杂因素并不总是一件容易的事情。这里的要点是，我们应当留意是否存在可能改变结论的混杂因素。让我们尝试几个例子。

阿拉斯加航空公司在五个存在竞争的主要机场，拥有优于另一家航空公司的准点运行记录，但其总体准点记录则不如竞争对手，为什么？因为阿拉斯加航空拥有许多飞往西雅图的航班，而西雅图的天气问题经常导致飞机延误。对于每个年龄群体，瑞典的女性死亡率都要低于哥斯达黎加，但瑞典拥有更高的女性总体死亡率，为什么？因为瑞典拥有更多的老年女性（老年人拥有相对较高的死亡率）。一项医学研究发现，一种手术对于小型和大型肾结石的治疗成功率均高于另一种手术，但其总体成功率却不如另一种手术，为什么？因为它经常被用于治疗大型肾结石（大型肾结石的治疗成功率相对较低）。

所有这些例子以及其他许多例子之所以存在辛普森悖论，是因为某种混杂因素对聚合数据产生了影响。不过，这并不意味着分解数据永远优于聚合数据。表 7.6 对两个假想的棒球选手进行了比较，将数据分解成了单日和双日。例如，科里（Cory）在双日的 100 次击球中击出 20 个安打，安打率为 20/100=0.200。在这些编造出来的数据中，两名选手恰巧在单日拥有更好的表现，科里恰巧在单日拥有更多的击球次数。因此，虽然吉米在单双日都具有更高的安打率，但是整个赛季安打率更高的人是科里。

表 7.6：谁是更好的击球手？

	双日	单日	所有日期
科里	20/100=0.200	90/300=0.300	110/400=0.275
吉米	61/300=0.203	31/100=0.310	92/400=0.230

根据这些数据，你认为谁是更好的击球手？我认为是科里，因为我们没有理由认为单双日是一个有意义的混杂因素。这只是数据中的一种巧合而已。如果我们根据每天的字母数分解数据，或者将日期按照字母顺序排列，我们可能会看到同样的现象。在这些情形中，分解数据中的模式仅仅是一种巧合，我们可以安全地将其忽略掉。聚合数据可以更加准确地衡量谁是更好的击球手。在其他一些情形中，混杂因素是真实的，忽略它们是一种危险做法。

我要再来一杯咖啡

咖啡拥有一个漫长而充满争议的历史。咖啡灌木几个世纪以前在非洲热带森林中被发现。最初，人们将磨碎的豆子和水果混在一起，连皮带瓤一齐吃下去。去除渣滓后饮用液态咖啡的做法似乎起源于 15 世纪的也门，随后传播到其他地区。在土耳其，咖啡成了日常生活的一个重要组成部分，人们甚至会对准新娘冲咖啡的能力进行评估。结婚以后，如果丈夫无法每天提供咖啡，妻子可以和他离婚。

咖啡目前是世界上最受欢迎的饮料之一。许多人喜爱咖啡，另一些人则认为咖啡具有成瘾性，对健康有害。咖啡被称为世界上使用最多的毒品。真相究竟如何？关于咖啡的第一项统计研究发生在 18 世纪的瑞典，那是一个有趣的实验。

咖啡在 17 世纪被引入瑞典并逐渐流行开来。不过，许多人将咖啡视为邪恶的饮品，认为它会使富人上瘾，使广大人民受到煽动。多年来，咖啡被课以重税，或者被完全禁止。后来，国王古斯塔夫三世（King Gustav III）开展了一项有趣的实验。据说，这是瑞典的第一项临床试验。

古斯塔夫是一位在许多方面表现开明的独裁者。他开展了许多经济和政治改革，赋予天主教徒和犹太人宗教自由，减少死刑和拷打，并且慷慨地支持艺术。不过，他的一些冒险行动则没有那么成功。为了团结人民，他曾向俄罗斯开战。瑞典皇家歌剧院的裁缝为一伙瑞典人制作了俄罗斯军装，这些人穿过俄罗斯边境，向一个瑞典边防哨所开火。以这场自导自演的挑衅事件为导火索，古斯塔夫对俄罗斯发动了一场“自卫”进攻。这是一场代价高昂的战争，双方失去了许多战士、船只、弹药和装备，却没有获得或失去任何土地。古斯塔夫国王没能利用共同的敌人将瑞典人民团结起来，反而导致了巨大的伤亡和经济代价，许多人对他丧失了信心。

古斯塔夫对酒精和咖啡的战争取得了同样失败的结果。为了筹集资金应对军事灾难，古斯塔夫将制造和销售酒精饮料的权利收归政府所有。这一政策的效果和美国的禁酒令类似。农民仍然在酿酒，酒精消费量不降反升。至于咖啡，一些著名的瑞典科学家认为咖啡是一种健康的滋补品，另一些人则持有相反的观点。古斯塔夫相信咖啡是一种毒药，并且决定证明这一点。他找到了两个犯下谋杀罪行，即将被斩首的男性双胞胎。古斯塔夫将他们的判决改为终身监禁，但是有一个条件。其中一个人需要每天喝三壶咖啡，另一个人需要每天喝三壶茶。两位由朝廷任命的医生将确保这些要求得到实施，并且需要在双胞胎去世时通知国王。古斯塔夫相信，他为那个喝咖啡的人准备了和斩首相同的命运——他的死将彻底证明咖啡是一种毒药。

结果，医生和古斯塔夫都死在了两个双胞胎的前面（古斯塔夫被人刺杀）。喝茶的兄弟最终在 83 岁那年去世，当时喝咖啡的兄弟仍然活得很健康。虽然这个结果出乎意料，但是瑞典政府仍然将咖啡的禁

令维持到了 19 世纪 20 年代。最后，瑞典政府终于决定完全允许瑞典人民去做他们想做的事情——饮用咖啡，饮用大量的咖啡。

值得注意的是，古斯塔夫在实验中明智地选择了两个完全相同的男性双胞胎，以消除性别、年龄和基因的混杂效应。最明显的缺点是，实验的样本太小了。如此小的样本无法得出任何具有统计说服力的结论——包括支持咖啡或者反对咖啡的结论。

多年以后，一些规模更大的研究认为古斯塔夫关于咖啡有害的观点是正确的。不过，这些研究都存在致命缺陷。一个反复出现的问题是，我们无法在现实中进行长期随机化实验。古斯塔夫是国王，可以用一对即将被斩首的双胞胎去做他想做的事情。我们无法让人们喝咖啡或者不喝咖啡。相反，我们只能使用观测性数据。我们观察到一些人喝咖啡；我们观察到另一些人不喝咖啡；我们对两个群体进行比较。问题是，选择喝咖啡的人和选择不喝咖啡的人之间可能存在系统性差异。

例如，1971 年的一项研究发现，同没有膀胱癌的人相比，患有膀胱癌的人更愿意喝咖啡——这意味着咖啡会导致膀胱癌。不过，这里存在一个混杂因素，那就是喝咖啡的人更愿意吸烟。导致膀胱癌的到底是咖啡还是香烟呢？1993 年，一项针对 35 项研究的严格分析报告证明了咖啡的清白，指出香烟才是真正的元凶。报告认为，“在扣除吸烟的影响以后，没有证据表明男性或女性的‘下尿道癌’风险出现了上升。”2001 年的一项研究证明了香烟会增加膀胱癌的风险，而咖啡则不具有这种效果。这项研究还提出了一个新的观点：同不喝咖啡的吸烟者相比，喝咖啡的吸烟者患上膀胱癌的可能性要小一些。咖啡似乎可以在一定程度上抵销香烟的致病效果。

下面是另一个例子。20 世纪 80 年代早期，备受尊重的研究员、哈佛大学公共卫生学院院长布赖恩·麦克马洪（Brian MacMahon）领导的一个团队发现，喝咖啡和胰腺癌之间存在紧密的联系。这项研究发表在世界顶级医学期刊之一《新英格兰医学期刊》上，并在全国范围内得到了报道。这个哈佛研究小组提出，如果人们停止喝咖啡，胰腺

癌发病率可能会得到极大的降低。麦克马洪遵守了自己提出的建议。在研究之前，他每天喝三杯咖啡。在研究之后，他戒掉了咖啡。

麦克马洪的研究对患有胰腺癌的住院病人以及患有其他疾病并被相同医生要求住院的病人进行了比较。这是一种针对观测性数据的便捷采样，因此是有问题的。这些医生中有许多胃肠专家。许多没有患癌的住院病人放弃了咖啡，因为他们担心咖啡会使溃疡和其他胃肠问题恶化。患有胰腺癌的病人并没有停止喝咖啡。这几乎可以保证胰腺癌患者之中拥有更多的咖啡饮用者。不是咖啡导致了胰腺癌，而是其他疾病导致其他人不再喝咖啡。

另一个问题是，麦克马洪之所以进行这项研究，是因为他认为胰腺癌可能与酒精或香烟有关。他考察了酒精，他考察了香烟，他考察了雪茄，他考察了烟斗。他没有发现任何关联。不过，他仍然在继续探索。他尝试了茶，他尝试了咖啡。终于，他发现了一个现象：胰腺癌患者喜欢饮用更多的咖啡。

当我们通过挖掘数据得到一种理论时，我们不能用搜刮过的数据检验这种理论。假设我抛 100 次硬币。在通过各种可能的角度考察数据以后，我注意到，在抛出两个正面和一个反面以后，下一个硬币是正面的可能性超过了一半。这种古怪的理论是通过仔细考察我自己的数据得到的，因此它当然可以被这些数据证实。要想进行公平的检验，我需要事先提出理论，然后用新的抛硬币实验对其进行检验。

随后的研究（包括麦克马洪的团队进行的一项研究）没能证实最初的结论："与之前的研究相反，我们没能观测到男性或女性的任何风险趋势。"美国癌症协会也同意这种观点："最近的科学研究发现，咖啡与胰腺癌、乳腺癌或者其他癌症风险之间没有任何关系。"随后的研究不仅没有证明麦克马洪最初的理论，反而发现喝咖啡似乎可以降低胰腺癌的风险——至少是对男性而言。

咖啡似乎具有许多健康益处，包括降低许多癌症、肝硬化、胆囊疾病、心血管疾病、阿尔茨海默病、帕金森病、斑块疾病以及痛风的风险。咖啡最大的风险似乎是伤害胃肠器官的保护层。正因为如此，

具有溃疡和其他胃肠问题的人才会被要求停止喝咖啡——这也是最初的胰腺癌研究存在缺陷的原因之一。

在这份存在缺陷的胰腺癌研究发表以后，2012 年，同一份著名期刊《新英格兰医学期刊》发表了迄今为止规模最大的一项研究，这项研究在 13 年时间里对 40 万人进行了跟踪。研究发现，扣除吸烟、喝酒和锻炼的混杂效应，同不喝咖啡的人相比，每天喝一杯咖啡的人在各个年龄段去世的概率都会下降 5 到 6 个百分点。对于每天喝两杯或三杯咖啡的人来说，男性的风险可以降低 10%，女性的风险可以降低 13%。

研究人员为什么不断改变想法呢？巧克力曾经是有害的，现在是有益的。葡萄酒曾经是有害的，现在是有益的。咖啡曾经是有害的，现在是有益的。阳光曾经是有害的，现在是有益的。好的变成了坏的，坏的变成了好的。这是因为，最初的研究存在缺陷——而这通常是因为人们忽略了混杂因素，或者为了寻找值得发表的结论而对数据进行了挖掘。

×如何轻松识破一本正经地胡说八道

如果一项研究支持你的观点，你会自然倾向于会意地点点头，认为你的观点得到了证实。更加明智的做法是进行仔细观察并考虑混杂因素。当一项研究看上去不合理时，你也应当采取这种做法。

例如，伯克利研究生录取政策歧视女性的说法看上去是合理的。不过，当人们开展深入调查，以确定问题最严重的院系时，他们发现了意想不到的现象——实际上，这些院系倾向于优待女性申请人。女性的总体录取率之所以偏低，是因为她们更喜欢申请录取率较低的教育计划。

由于潜在的自选择偏差和混杂因素，观测性研究存在固有的挑战性。应时刻对利用数据发现理论的研究保持警惕。

第 8 章 状态火热的雷·阿伦

在宿敌波士顿凯尔特人和洛杉矶湖人之间进行的美职篮（NBA）总决赛第二场比赛中，凯尔特人后卫雷·阿伦（Ray Allen）连续投中了七个三分球。一名队友说，这“令人难以置信”。另一名队友认为这件事“不可思议”。一位体育新闻撰稿人写道，“阿伦找到状态了。”另一个人写道，阿伦“进入了仅仅属于现实版超级明星和电影角色的投篮区域”。

雷·阿伦的情况并不是个例。连续得手和连续失误在体育运动中很常见。许多篮球选手连续多次投篮得分（或者投篮不进）。许多橄榄球四分卫连续多次传球成功（或者传球失误）。许多棒球选手连续多次安打（或者出局）。球迷和选手看到了这种连续现象，认为选手状态很好或者状态糟糕。珀维斯·肖特（Purvis Short）在 12 年的美职篮生涯中平均每场比赛得到 17 分，他还在一场比赛中得到了 59 分。肖特说出了人们的普遍看法：“你处在仅仅属于你自己的世界里。这很难描述。不过，篮筐似乎变得很宽。你知道，不管你做什么，球都会进入篮筐。”

和以前一样，我们仍然是在看到一种模式以后编造出符合这种模式的理论。如果篮球选手连续多次投篮得分，这一定是因为他的状态很好，投篮得分的概率出现了上升。如果选手连续多次投篮不中，这一定是因为他的状态不好，投篮得分的概率出现了下降。大多数球迷

和选手并没有认识到，即使每次投篮、传球或击球与之前的投篮、传球或击球没有关系，也会出现完全来自巧合的连续现象。

例如，假设我抛 20 次硬币，每次抛硬币代表一次投篮。正面代表命中，背面代表不中。一连串正面代表状态火热，一连串背面代表状态不佳。下面是我的结果：

正面 正面 正面 背面 背面 背面 背面 正面 正面 背面
正面 正面 正面 正面 背面 正面 正面 背面 背面 正面

从第 4 次开始，连续出现了 4 个背面。从第 11 次开始，连续出现了 4 个正面。我经历了状态不好的时刻，随后又经历了状态很好的时刻，尽管每次抛硬币都是独立的！

我将这个实验重复了 10 次。在 7 次实验中，我抛出了连续至少 4 个正面或背面。我两次抛出连续 4 个正面或背面，4 次抛出连续 5 个正面或背面，一次抛出连续 10 个正面或背面。表 8.1 显示了抛 20 次硬币时出现各种连续次数的理论概率。连续抛出 4 个或更多正面或背面的概率是 0.768。在我的 10 次实验中，有 7 次出现了这样的情况。我所抛出的连续 10 个正面或背面不太常见，但如果像我这样将实验重复 10 次，这种情况也有大约 11% 的概率。

表 8.1：抛 20 次硬币连续出现正面或背面的概率

连续次数	恰好是这个连续次数的概率	至少是这个连续次数的概率
<3	0.021	1.000
3	0.211	0.979
4	0.310	0.768

5	0.222	0.458
6	0.121	0.236
7	0.061	0.115
8	0.029	0.054
9	0.013	0.025
>9	0.012	0.012

当然，这是抛硬币，不是真正的投篮。不过，这正是问题所在！在完全随机的抛硬币中，常常会出现仅仅来自巧合的连续命中或不中。这并不能证明运动员的优秀状态或糟糕状态仅仅是一种巧合，但它却在提醒我们，优秀的状态无法确保连续成功，糟糕的状态也不能保证连续失败。优秀或糟糕的状态也许仅仅是运气而已。

雷·阿伦在美职篮职业生涯中投出了超过 7000 个三分球，命中率为 40%。想象一个拥有 40% 正面概率的硬币被抛了 7000 次。实际上，在这 7000 次里面，在某个地方，几乎一定会出现连续 7 个正面。雷·阿伦的连续命中也许并不比连续 7 个正面具有更多的含义。

小数定律

丹尼尔·卡尼曼（Daniel Kahneman）和阿莫斯·特沃斯基（Amos Tversky）合作发表了许多论文，包括发现我们的判断如何受系统性偏差和误差影响的开创性研究。2002 年，卡尼曼获得了诺贝尔经济学奖。特沃斯基无法分享这个奖项，因为他已经过世了。不过，卡尼曼表示："我感觉这是一个联合奖项。我们共事了十几年时间。"

卡尼曼和特沃斯基观察到的认知错误之一是相信小数定律。想象我们从一个装满红球和蓝球的巨大容器中取出 10 个球。小数定律指的是这样一种错误观念：如果容器中 50% 的球是红球，那么我们取出的 10 个球中将有 5 个红球。事实并非如此。我们取出 5 个红球和 5 个蓝球的概率只有大约 25%。大多数时候，红球和蓝球的数量是不等的。

错误的小数定律将导致两个相互关联的错误。第一个错误叫做赌徒谬误。如果我们取出的前 3 个球是红球，那么我们倾向于（错误地）认为下一个球很可能是蓝球，因为我们最终一定会得到 5 个红球和 5 个蓝球。类似地，如果一个均匀的硬币被抛掷 10 次，前 3 次都是正面，那么下一次很有可能是背面，因为我们最终将会得到 5 个正面和 5 个背面。第 10 章将会更加详细地讨论这种谬误。

第二个错误发生在我们不知道容器中有多少红球和蓝球的情形中。如果我们取出 5 个球，其中 4 个球是红球，我们就会（错误地）认为容器中一定有 80% 的球是红球。因此，下一个球有 80% 的可能性是红球。

类似地，如果篮球选手在 5 次投篮中 4 次命中，我们可能认为他下次投篮命中的概率是 80%。如果这个选手在 5 次投篮中 4 次不中，那么他下次投篮命中的概率只有 20%。根据很小的投篮样本，我们认为选手的状态由好转坏，从 80% 的命中率转变成了 20% 的命中率。我们没有意识到，即使选手每次投篮命中的可能性都是 50%，他也会时而五投四中，时而五投一中。这种现象无法说明任何问题。

我们天生倾向于寻找模式并且相信我们看到的模式背后一定存在某种合理的解释。上述错误观念就是这种倾向的一个例子。1937 年和 1938 年，真力时广播公司广播了一系列每周超感知觉实验。这个广播节目邀请听众猜测播音室里的“发送者”正在观看的符号，并把猜测结果寄给电台。这种实验涉及五个随机“二选一”（比如圆圈或者方块），类似于 5 次抛硬币。

这项在全国范围内进行的实验并没有证明超感知觉的真实性。相反，实验的主要结论是，人们低估了随机数据中模式的出现频率。听

众选择了看上去具有随机性的序列，避开了看上去不具有随机性的序列。例如，在一项实验中，121 位听众选择了下面这个序列：

□□○□○

只有 35 个人选择了

□○□○□

只有 1 个人选择了

□□□□□

如果方块和圆圈是随机选择的，就像抛 5 次硬币一样，那么上面这些序列拥有完全相同的出现概率。不过，听众感觉方块和圆圈应当得到平衡（第 10 章将会讨论这种错误的平均定律），因此不会出现全是方块或者全是圆圈的情况。他们还认为方块和圆圈完美交替式的平衡不太可能出现，因为它看上去不具有随机性。听众不愿意猜测连续五个方块或者方块和圆圈完美交替的序列，因为他们不相信这些序列会随机出现。

现在从另一个方向考虑这个问题。你不需要猜测圆圈和方块序列的结果。相反，考虑你看到其中某种模式时的反应。第一个模式没有任何独特之处。第二个模式比较奇特，第三个模式更加奇特。虽然我们知道圆圈和方块出现的可能性是相等的，但是我们仍然会低估连续 5 个方块或者完美交替序列完全随机出现的可能性。相反，我们认为这些模式背后一定存在某种特殊的因素。也许，电台人员对圆圈和方块的选择不是随机的？

同样的道理，当一个平时拥有 50% 命中率的篮球选手连续 5 次投篮命中时，我们认为一定存在某种独特的因素。选手状态很好，下次投篮很可能命中。连续 5 次传球成功的四分卫状态很好，下次传球很可能成功。连续 5 次抓到好牌的牌手运气很好，下一局很可能抓到好牌。

我们不断低估巧合在生活中的存在性，没能认识到随机性会生成看上去有意义但实际上毫无意义的模式。我们很容易被那些对无法解释的事情做出解释的说法所引诱。

一项篮球研究

三位杰出的心理学家——康奈尔大学的托马斯·吉洛维奇（Thomas Gilovich）以及斯坦福大学的罗伯特·瓦隆（Robert Vallone）和阿莫斯·特沃斯基——对“手顺”现象做了一项有趣的研究。他们调查了 100 名篮球迷，发现 91% 的人相信选手“在两三次投篮命中以后投篮命中的可能性高于两三次投篮不中以后投篮命中的可能性。”他们还要求这些球迷估计一个假想的、拥有 50% 命中率的选手在上次投篮命中以及不中的情况下投篮命中的概率。球迷的平均估计值分别是 61% 和 42%。

篮球选手也相信状态火热和糟糕的说法。在他们调查的 7 位职业选手中，5 个人（71%）相信选手在之前两三次投篮命中以后投篮命中的可能性要高一些。对于拥有 50% 命中率的假想选手在上次投篮命中以及不中的情况下投篮命中的可能性，职业选手的平均估计值分别是 62.5% 和 49.5%。

吉洛维奇、瓦隆和特沃斯基考察了各种篮球数据，认为这种常见的感受是错误的。他们最有说服力的数据来自美职篮费城 76 人队 1980～1981 赛季的表现。这些教授对每名选手连续 2 次或 3 次投篮命中以后投篮命中的频率与这名选手连续 2 次或 3 次投篮不中以后投篮命中的频率进行了比较。他们发现，事实上，选手在连续投篮命中以后的整体表现稍差于连续投篮不中以后的表现。

这种分析的一个问题是，上述数据没有考虑两次投篮的时间间隔。一名选手的连续 2 次投篮可能间隔 30 秒，间隔 5 分钟，位于一场比赛的上下半场，甚至出现在不同的比赛中。球迷和选手并不认为一

个人周二在费城的表现会影响他周四在波士顿的投篮命中概率。另一个问题是，连续命中的选手可能倾向于做出难度更大的投篮动作。这可以解释选手在投篮命中以后整体表现更加糟糕的原因。

此外，当一名选手被认为状态火热或糟糕时，对方可能会对他采取不同的防守策略。投篮选择可能还会受到分数、比赛剩余时间以及双方选手累计犯规次数的影响。这个问题存在许多混杂因素。

美职篮首发队员平均每场比赛进行 10 ~ 20 次投篮，每个半场 5 ~ 10 次。如果“手感好”是一种相对温和的现象，那么基于 5 ~ 10 次投篮的统计检验不太可能发现这种现象，尤其是当数据中隐藏着不同投篮难度、不同投篮间隔以及其他混杂影响因素时。

吉洛维奇、瓦隆和特沃斯基的结论不仅与球迷和选手的观念相抵触，而且与大量证据相矛盾。在大量体育运动中，当运动员充满自信时，他们可以做出更好的表现。在一项掰手腕研究中，竞争者被事先告知他们比对手强或者比对手弱。当对决双方获得错误信息、认为弱者更强时，较弱的一方在 12 场比赛中赢了 10 场。当对决双方获得关于谁是强者的正确信息时，较强的一方赢得了所有 12 场比赛。

一种看上去比较合理的观点是，连续投篮命中的篮球选手将会变得更加自信，这种信心可以帮助他做出更好的表现。也许，好手感是真实的，但是由于投篮选择、投篮间隔以及防守调整等混杂因素而很难在篮球比赛中被检测出来。

那么，其他运动项目的证据呢？高尔夫球和掷飞镖领域有一些关于状态火热和状态糟糕的实验证据。不过，在这些研究中，志愿者需要反复打高尔夫球和掷飞镖，但他们的报酬却很微薄。（你相信每天 5 美元的价格吗？）面对大量试验，他们几乎没有理由认真对待这件事情，因此研究人员观测到的火热和糟糕状态可能源自动力不足的志愿者专注程度的起伏。如果志愿者专心实验时的成功率较高，感到厌倦时的成功率较低，命中和失误就会在数据中出现聚集，形成连续成功和连续失败的现象。在这些连续成功和连续失败之中，他们的能力并没有出现波动，真正出现波动的是他们的专注力。

为了绕过篮球研究和人工实验中存在的问题，我开始寻找高水平运动员在没有混杂因素的情况下为了真正的利益而全力竞争的运动项目。选手的每次尝试最好具有相同的条件和相同的难度，两次尝试之间最好间隔很短的时间。我想到了两个运动项目：职业掷马蹄铁和保龄球。它们并不是非常吸引人的运动，但它们的确具备我所需要的特点。

小沃尔特·雷·威廉姆斯

一位名叫小沃尔特·雷·威廉姆斯的优秀运动员恰好在两个运动项目上都有着不俗的表现。威廉姆斯最初用右手投掷马蹄铁。10 岁那年，他参加了一场锦标赛，在 50 次投掷中投出了 45 个套环，获得了“神投手”的外号。13 岁那年，他伤到了右手手指，并且开始学习左手投掷。18 岁那年，他伤到了右手手腕，因此改到了左手。他的右手套环率是 85%，左手套环率是 50%。是的，你没听错；他的左手比我们一般人的右手还要厉害。

威廉姆斯用右手获得了 6 个世界冠军。当他的套环率下降到 70% 时，他转换到了左手。46 岁那年，他用左手在世界锦标赛中获得了第二名。威廉姆斯说，现在，他的两只手在这个项目中的表现一样好，但他更喜欢使用左手。

他在保龄球领域的表现也许更加出色，而且这显然为他带来了更加丰厚的财务回报。（你无法以投掷马蹄铁为生，但你可以通过投掷保龄球成为百万富翁。）威廉姆斯 7 次成为职业保龄球员协会（PBA）年度选手，在 PBA 锦标赛中赢得了将近 500 万美元奖金。不过，他最喜爱的仍然是马蹄铁。他开着一辆房车周游全国，从一个马蹄铁或保龄球赛场赶到另一个赛场，并且尽量挤出一些时间去打高尔夫——他有 3 个差点（handicap）。在被问及为什么能够做到这些

时，他的回答很简单：“长期以来，我一直是这样做的，我觉得这很正常。”

威廉姆斯有许多连续多次命中的经历。1991 年，他以 32 岁的年龄连续投出了 56 个马蹄铁套环。他在 PBA 比赛中打出了 8 个完美球局（连续 12 次全中）。不过，他当时已经获得了两个体育项目的世界冠军，并且投掷了许多马蹄铁和保龄球。也许，这些连续命中仅仅是幸运的巧合。

威廉姆斯是一个聪明的运动员。他曾回忆自己高中时在当地大学参加计算机课程时的情景：“最初，我和许多对计算机一无所知的人一样，感到有点害怕，因为我们觉得计算机很聪明。实际上，计算机是很笨的。它们只会按照你的要求工作。如果你让它们去做错误的事情，它们就会去做错误的事情。”我说过，他的确是一个聪明的运动员。

威廉姆斯在加州大学波莫纳分校主修物理学，辅修数学。在精彩的保龄球纪录片《一群普通绅士》中，威廉姆斯表示，如果保龄球这条路走不通，他“很可能会成为教师，或者为美国宇航局工作。”在加州大学波莫纳分校的毕业论文中，他写了一个预测保龄球轨迹的计算机程序。他还写了一个软件程序，用于记录马蹄铁锦标赛的详细结果。当我寻找马蹄铁比赛数据中，我无意中发现了威廉姆斯及其收集的数据。他对“好手感”现象也很感兴趣，而且愿意和我分享他的数据。

投掷马蹄铁

马蹄铁比赛场地有两个相距 12 米的标桩。在每局比赛中，每位选手向同一个标桩投掷两个马蹄铁。所有 4 个马蹄铁投掷完毕后开始计分。套住标桩的马蹄铁叫做“套环”，记三分。位于标桩 15 厘米范围内

的非套环记一分。在常规抵销计分法中，两位选手投出的套环相互抵销，与标桩距离相等的马蹄铁也相互抵销。如果一位选手投出两个套环，对手投出一个套环，第一位选手得三分。如果两位选手都投出两个套环，双方不得分。第一位达到 40 分的选手将成为获胜者。

在世界锦标赛中，入围的 16 位选手捉对厮杀，每位选手需要与其他 15 位选手依次过招。最终排名是由选手的总体胜负记录决定的。

在顶级锦标赛中，选手通常投出 60% 到 80% 的套环，比赛通常持续二三十局。历史上最伟大的比赛之一出现在 1965 年的世界锦标赛上：在持续两个半小时、长达 97 局的马拉松式较量中，套环率 89.7% 的雷·马丁（Ray Martin）输给了套环率 90.2% 的格伦·"雷德"·亨顿（Glen "Red" Henton）。

这里的"好手感"问题是，选手在一局比赛中投出的套环数量是否取决于之前几局投出的套环数量。由于世界级投手通常不会出现双误，因此我将每个选手每局的表现描述成双套环或非双套环。顶级选手投出双套环的次数大约是一半，这与抛硬币非常类似。

表 8.2 使用了威廉姆斯向我提供的 2000 年和 2001 年世界锦标赛的数据。根据这张表，选手在双套环之后投出双套环的可能性高于非双套环之后投出双套环的可能性（0.548 对 0.501）。他们在连续两次双套环之后投出双套环的可能性高于两次非双套环之后投出双套环的可能性（0.545 对 0.490）。这些"好手感"模式意味着存在另一种相同的"差手感"模式：非双套环之后出现非双套环的可能性高于双套环之后出现非双套环的可能性（0.499 对 0.452）。

表 8.2 中的经验性差异看上去可能不太明显，但它足以决定谁是冠军，谁是失败者。

表 8.2：双套环或非双套环之后双套环或非双套环的频率

	双套环频率	非双套环频率

两次双套环之后	0.545	0.455
一次双套环之后	0.548	0.452
一次非双套环之后	0.501	0.499
两次非双套环之后	0.490	0.510

在 64 名选手中，51 名选手在双套环之后投出双套环的可能性高于非双套环之后投出双套环的可能性。在 63 名选手中，48 名选手在两次双套环之后投出双套环的可能性高于两次非双套环之后投出双套环的可能性（一名选手的两种可能性相等）。这些差异均具有很高的统计显著性。投掷马蹄铁的随机性不像抛硬币那么大。选手的确会获得好手感和差手感。

保龄球

在保龄球运动中，10 个目标瓶被排列成金字塔形，与球道起始处相距 18 米。每局比赛包含 10 格。如果选手在某一格第一次投球时击倒所有 10 个球瓶，这一格将被记作全中。如果选手没有投出全中，倒下的瓶子将被清走，选手将获得击倒剩余球瓶的第二次机会。如果选手在第二次投球时击倒剩余球瓶，这一格将被记作补中。

选手每格的基本分等于被击倒的球瓶数。如果是补中，得分为 10 分加上下次投球时被击倒的球瓶数。如果是全中，得分为 10 分加上下两次投球时被击倒的球瓶数。如果第十格投出补中或全中，选手将获得奖励投球机会（补中奖励 1 次，全中奖励 2 次）。完美球局包含 12

次全中（十格加上两次奖励投球），得分为 300 分。职业保龄球选手投出全中的概率为 60% 左右，平均每局得分超过 200 分。

在 PBA 锦标赛中，所有选手先投 9 局。9 局得分排在前 64 位的选手进入下一轮，然后再投 9 局。前 32 名选手进入对抗赛阶段，选手将根据前 18 局的表现获得种子编号并捉对厮杀。1 号种子对 32 号种子，2 号种子对 31 号种子，依此类推。

表 8.3 使用了 2002 ~ 2003 赛季 PBA 对抗赛阶段的数据。根据这张表，选手在前一格投出全中以后投出全中的可能性高于前一格投出非全中以后投出全中的可能性（0.571 对 0.560）。他们在前两格投出全中以后投出全中的可能性高于投出两个非全中以后投出全中的可能性（0.582 对 0.546）。

和投掷马蹄铁一样，表 8.3 中的差异看上去可能不太明显，但它足以决定谁是胜利者，谁是失败者。

表 8.3：全中或非全中之后全中的频率

	全中的频率	非全中的频率
两次全中后	0.582	0.418
一次全中后	0.571	0.429
一次非全中后	0.560	0.440
两次非全中后	0.546	0.454

在 134 名保龄球选手中，80 名选手全中后投出全中的可能性高于非全中后投出全中的可能性。而且，在 110 名选手中，77 名选手在两次全中后投出全中的可能性高于两次非全中后投出全中的可能性（24

名选手数据不足，无法进行有效比较）。和投掷马蹄铁一样，这些保龄球差异具有很高的统计显著性。

我们也许可以将连续 12 次全中组成的完美球局视为终极好手感。在 2002～2003 赛季 PBA 巡回赛的对抗赛阶段，出现了 19 场完美球局。如果“好手感”只是一种传说，那么完美球局的预期场次只有这个数字的一半左右。42 名选手在前 10 格中全部投出了全中，需要加投两次。根据这项数据，如果“好手感”只是一种传说，那么完美球局的预期场次也只有 19 场的一半左右。这两个结果均具有统计显著性。

我们的结论比较复杂。我们被模式诱惑，希望获得这些模式的解释。当我们看到一连串成功时，我们认为“好手感”提高了成功的可能性。当我们看到一连串失败时，我们认为“差手感”提高了失败的可能性。对于抛硬币，我们很容易驳斥这种理论，但在涉及人类时，情况就不同了。我们显然会产生或好或坏的情绪，这将导致我们的能力出现起伏。问题是，这些波动是否重要，是否可以忽略。

许多关于好手感的实证研究并没有给出具有说服力的答案，因为这些研究的数据中存在混杂因素。例如，投篮涉及不同的投球位置和不同的防守压力。所以，我们很难判断导致选手表现起伏的原因是他的能力波动，还是他的不同的投篮方式或者对手的不同防守策略。

不具有这些混杂因素的马蹄铁和保龄球数据显示出了好手感和差手感的证据。我们所观测到的表现波动足以区分胜利和失败——但它并没有人们想象的那么明显。

×如何轻松识破一本正经地胡说八道

我们喜欢在数据中寻找模式并为我们所看到的模式编造一些理由，这是无法避免的事情。因此，我们很容易相信好手感和差手感的说法是真实的，相信成功率会出现极大的波动。记住，即使在随机的抛硬币实验中，也会出现仅仅来自巧合的、引人注目的连续成功和连续失败现象。

好手感和差手感很可能的确存在，但它比我们想象的要小得多。

第 9 章
胜者的诅咒

西北大学经济学教授贺拉斯·西克里斯特（Horace Secrist）职业生涯可谓相当杰出。他写了 13 本教材，曾担任西北大学经济研究局局长。1933 年，在后来被称为“大萧条”的巨大国家经济灾难中，他出版了一本书，希望能够解释灾难的原因，提供解决方案，维护自己的学术名誉。结果，这本书反而败坏了他的名声。

当时的传统经济观点认为，从整个国家来看，供给会创造出与之相应的需求。像总需求不足这样的事情是不可能存在的，因为企业不仅雇用工人提供商品和服务，而且为工人支付工资，以确保一个国家的公民有能力购买这些商品和服务。像失业率这样的问题是不可能存在的，因为劳动力市场可以平衡需求和供给，确保每个想要工作的人都能找到工作。这就是需求供给平衡的含义：公司能够招到他们想要招聘的人数，想要工作的人能够找到工作。

英国经济学家约翰·梅纳德·凯恩斯（John Maynard Keynes）对全球经济大萧条进行了研究，认为传统观念是错误的。从 1929 年到 1933 年，美国的产量下降了一半，失业率从 3% 上升至 25%。农场收入下降了 1/3，超过十万家企业破产，包括全国超过 1/3 的银行。绝望的人们在街角乞讨，在垃圾堆里翻找食物。在英国，总产量下降了 1/3，失业率达到了 20%。

凯恩斯总结道：

> 马尔萨斯之后的职业经济学家显然对其理论结果与事实观察缺乏一致性的现象无动于衷……这使经济学家们被视为赣第德式的人物，他们脱离这个世界，专心浇灌自己的菜园，他们认为只要我们不加干预，整个世界就会在所有可能的方向中朝着最美好的方向发展……经典理论也许代表了我们的经济应当具有的状态。不过，如果认为现实与理论相符，我们就会忽视眼前的困难。

凯恩斯认为，需求不足不仅是可能的，而且是大萧条的原因。一个人只要睁开眼睛，就会看到这一点。他认为，供给不会自动创造需求；相反，需求常常会创造供给。如果人们减少对家具的开销，企业就会减少家具产量，解雇他们所聘用的一些家具制造者。这些失去工作和收入的家具制造者将会削减食品、服装和娱乐开销，导致这些行业的公司减少产量和员工。这种效应将波及整个经济，因为每个人的开支为其他人提供了收入，每一次开支的降低都会减少其他人的收入和消费能力。失业人数可能会越来越多，并且可能维持很长时间。

1936 年，凯恩斯发表了经典论文《就业、利息和货币通论》，颠覆了经济学理论。实际上，这篇论文创造出了一个全新的经济学分支，叫做宏观经济学。

在此之前，贺拉斯·西克里斯特已在学术领域浸淫多年，深受古典经济学的束缚。西克里斯特认为，大萧条的原因一定出现在供给侧。因此，他开始研究企业实践。西克里斯特及其助手花了 10 年时间，收集和分析了美国百货商店、服装店、五金店、铁路和银行等 73 个不同行业的数据。他整理了从 1920 年到 1930 年多项企业成功指标的年度数据，包括利润对销售额的比率、利润对资产的比率、费用对销售额的比率、费用对资产的比率。对于每个比率，他根据 1920 年的比率值将每个行业的公司分成了四个部分：最优的 25%、次优的 25%、第三好的 25% 以及垫底的 25%。接着，他计算了前 1/4 公司从 1920 年到 1930 年每年的平均比率。他对其他三个部分的公司进行了同样的计

算。每个部分的比率值都在随时间的推移而收敛。1920 年位于前两个部分的公司在 1930 年更加接近平均水平。1920 年位于后两个部分的公司在 1930 年更加接近平均水平。最极端的两个部分表现出了更大的变化。

显然，他发现了一个普遍成立的经济真理。随着时间的推移，最成功和最不成功的公司倾向于更加接近平均水平。美国企业正在向平庸靠拢。他在《平庸在商业中的胜利》一书中记录和描述了这种理论。这本书是一部统计学力作，共 468 页，包含 140 张表格和 103 张图表，有力地支持了他那不同寻常的结论。

西克里斯特的结论是:

> 完全自由的行业准入和持续竞争意味着平庸的持续。新的公司是由相对“不胜任”，至少是缺乏经验的人成立的。如果一些公司取得成功，它们必须满足它们所归属的行业和市场的竞争要求。不过，出色的判断、经营意识和诚信总是无法抵抗肆无忌惮、愚蠢轻率、信息不畅和考虑不周的做法。结果是，零售行业过于拥挤，店铺很小而且缺乏效率，大量企业能力不足，费用相对高昂，利润微薄。只要某个领域可以自由进入（事实的确如此），只要竞争是“自由”的（在上述范围内，事实的确如此），那么优势和劣势往往无法持续。相反，平庸往往大行其道。

美国的经济问题显然源于他所发现的新的经济原则：竞争压力将不可避免地削弱优秀企业的才能。显而易见的解决方案呢？保护优秀公司免于希望进入这个市场的不太胜任的公司所带来的竞争。

在出版这部作品之前，西克里斯特向 38 位杰出的统计学家和经济学家寻求意见和批评。他们显然没有发现任何问题。作品出版以后，著名同行们最初发表的评论充满了溢美之辞。

> 这本书出色地说明了如何利用统计研究将经济学理论转变成经济学定律，将一门定性科学转变成定量科学……这本书以极为可信的方式展示了一位有能力的统计学家兼经济学家如何以刻苦的、长期持续的、有思想的、极为成功的努力加强我们关于竞争事实和理论的知识。（《政治经济学期刊》）
>
> 你将情不自禁地赞美作者及其助手在完成一项极为艰苦的任务时表现出来的热情和执着。（《皇家统计学会期刊》）
>
> 作者总结说，相互依赖的商业结构中竞争力量的相互作用将确保“平庸的胜利”。作者的研究方法具有充分的科学性。（《美国经济评论》）
>
> 从社会角度看，具有竞争性商业制度的经济体未来毫无希望，只有两点例外：这些研究可能会导致一些商业规划被引入；尽管优秀单位倾向于沦为平庸，但是这种平庸的整体水平可能会通过某种形式的私人或社会控制得到提升……这些结果为商业人士和经济学家带来了一个具有某种悲剧性色彩的持续问题。（《美国政治和社会科学院年报》）

杰出统计学家哈罗德·霍特林（Harold Hotelling）写了一篇犀利的评论，礼貌而坚定地说明了西克里斯特10年的工作成果无法证明任何事情。西克里斯特不但没有维护自己的名誉，反而成了被均值回归愚弄的一个经典案例。

均值回归

我曾经和一个学生对棒球领域的均值回归做了一些研究。我们没有写出 468 页的著作，但我们至少在看到均值回归时认识到了这一点。发表这篇论文的期刊评论员写道：

> 几乎没有哪个统计事实比均值回归更加有趣，原因有两点。首先，人们几乎每天都会在生活中遇到它。其次，几乎没有人理解这种现象。
>
> 这两个原因的叠加使均值回归成了人类决策最基本的错误来源之一，在医学、教育、政府甚至体育领域导致了许多错误的推理。

没错，我们几乎每天都会遇到均值回归现象，而且几乎没有人理解这种现象。就像西克里斯特的例子说明的那样，这是一种致命的组合。

要想理解回归，假设 100 个人被问到 20 个关于世界历史的问题。每个人的“能力值”是他在大量此类测试中的平均得分。一些人的能力值是 90，一些人的能力值是 80，一些人的能力值近乎为零。

一个能力值为 80 的人在测试中拥有 80% 的平均正确率，但他不会在每次测试中得到 80% 的正确率。想象一个拥有无数问题的测试库。根据幸运程度，一个能力值为 80 的人在一项测试中知道超过 80% 问题的答案，在另一项测试中知道不到 80% 问题的答案。一个人在任何一项测试中的得分都是对其能力值的不完美测量。

我们能够根据一个人的测试分数做出哪些推断呢？一个重要推断是，一个测试分数相对其他人比较高的人很可能得到了高于自身能力的分数。在一项测试中得分在 90 ~ 95 之间的人可能具有更加普通的能力值（比如 85 ~ 90、80 ~ 85 或者 75 ~ 80），只是这一次发挥出了超常水平；他也可能具有更高的能力值（比如 95 ~ 100），只是这一次

表现不佳。前一种情况的可能性比较大，因为能力值低于 90 ~ 95 的人比能力值高于 90 ~ 95 的人要多。

如果这个人的能力实际上低于 90 ~ 95，那么当他进行另一次测试时，他的分数很可能也会低于 90 ~ 95。类似地，得分远低于平均水平的人可能表现不佳，并且可能在随后的测试中得到高一些的得分。这种“得分远离均值的人在第二次测试中倾向于获得更加接近均值的得分”的现象是均值回归的一个例子。

我们可以在许多环境中遇到回归现象——尤其是当我们想要测量的某件事情得到不完美的测量结果时。标准化考试显然是对能力的不完美测量。因此，由于较低的考试成绩而获得特殊辅导的学生通常并不像他们的分数显示的那样糟糕。所以，即使辅导老师什么也不做，只是打着响指说“给我提高！”，这些学生也有可能在随后的考试中做出更好的表现。

医学检测是对患者情况的不完美测量，因此检测结果会出现回归现象。假设一名患者在常规医学检查中接受 20 项检测，其中一项检测结果令人担忧。患者的状况很可能不像检测结果显示的那样糟糕。在接受一些治疗以后，即使这种治疗毫无价值，患者也可能出现改善。

投资成功是对投资者才能的一种不完美测量。因此，我们也会看到回归现象：平均来看，在任何一年做出最佳股票选择的投资咨询师都会在第二年变得更加平庸。

第 7 章对不同网页布局（比如一次点击和二次点击形式以及不同的页面颜色和字体）进行测试的互联网公司也会出现回归现象。公司根据下列实验提出建议。当一名用户访问网站时，网站用随机事件生成器将用户带到多个页面版本中的某个版本。接着，公司记录用户的反应（用户在页面上的停留时间；用户是否一层层点击下去，以获得更多信息等等）。经过几天测试，公司需要汇报哪个页面版本最为成功。该公司反复遇到的一个问题是，当客户听从公司的建议并更改布局时，他们所获得的实际收益通常低于实验中观测到的收益。你如何解释这种现象呢？

当然是均值回归啦！每个实验得分都是对布局收益“能力”的不完美测量。由于实验中存在随机性（这当然是事实），得分最高的布局的未来收益很可能比实验结果更加接近均值。如果实验中用户点击量出现了 30% 的增长，实际增长率可能只有 20%。这种布局没有任何问题。它仍然是最成功的布局。只不过，公司需要认识到，“实际收益比实验预测值更加温和”是一种极为自然的现象。

诺贝尔奖获得者丹尼尔·卡尼曼曾告诉以色列飞行教官，如果新兵接受表扬而不是惩罚，他们可以实现更快的进步。一位高级教官提出了强烈的反对意见：

> 我曾在许多场合对漂亮地执行某种特技动作的飞行学员提出表扬。总体而言，当他们再次训练时，他们的表现会退步。另一方面，我经常对表现糟糕的学员大吵大嚷。总体而言，他们下次会做出更好的表现。所以，请不要告诉我们鼓励是有效的，惩罚是无效的，因为实际情况恰恰相反。

卡尼曼立即意识到，这位教官受到了均值回归的欺骗。做出最佳飞行表现的学员通常不具有他们表现出来的那种远高于平均水平的能力。平均来说，不管教官表扬他们，叱责他们，还是一言不发，他们都不会在下次飞行中做出同样好的表现。指责卡尼曼的高级教官错误地认为他的表扬导致学员表现退步；实际上，这些学员并没有看上去那样优秀。类似地，平均来说，飞行表现最为糟糕的学员并没有看上去那么无能；如果教官能够控制住自己的吼叫，这些学员可以在下次飞行中做出更好的表现。

卡尼曼后来写道：

> 这是一个快乐的时刻，因为我理解了这个世界的一个重要真理：由于我们倾向于在其他人表现出色时奖励他们，在

> 其他人表现糟糕时惩罚他们，又由于均值回归现象，因此从统计上看，我们将由于奖励别人而受到惩罚，由于惩罚别人而受到奖励，这是人类社会的一个组成部分。

卡尼曼不是第一个认识到均值回归重要性的人。在 19 世纪，弗朗西斯·高尔顿爵士（Sir Francis Galton）在研究父母及其成年孩子身高时观察到了均值回归现象。身材异常高大的父母倾向于拥有矮一些的孩子，身材异常矮小的父母倾向于拥有高一些的孩子。不过，高尔顿得出了错误的结论：人类身高正在向平庸回归；实际上，高尔顿这项研究的题目是《世代身高向平庸回归》。

均值回归并不意味着每个人很快都会得到相同的历史测试分数，它也不意味着每个人很快都会具有相同的身高。均值回归意味着观测到的身高是对我们遗传自父母并传给孩子的基因因素的不完美测量。一个身高 2 米的人可能拥有 1.8 米的基因并受到了积极的环境影响，或者拥有 2.1 米的基因并受到了消极的环境影响。前一种情况的可能性比较大，因为拥有 1.8 米基因的人比拥有 2.1 米基因的人多得多。因此，我们所观测到的非常高的父母的身高通常高于他们的基因身高以及他们孩子的平均身高。

回归是双方向的，因为它仅仅反映了随机波动。身高很高的父母倾向于拥有矮一些的孩子，身高很高的孩子倾向于拥有矮一些的父母。我的身高是 1.9 米，我不仅比我的成年孩子高，而且比我的父母高。

这种现象不限于身高。回归存在于无法通过观测型特点准确反映出来的任何遗传特性之中：身高、体重、智力、足部尺码、头发密度。异常的父母通常拥有不太异常的孩子，异常的孩子通常拥有不太异常的父母。

西克里斯特的愚蠢

同样的道理，西克里斯特关于成功和不成功公司的研究也涉及均值回归。在任何给定的年份，最成功的公司可能拥有更多的好运，取得不仅优于其他公司，而且优于自身长期利润率的表现。最不成功的公司则恰恰相反。这就是顶尖和垫底公司随后的表现通常更加接近公司平均水平的原因。与此同时，它们的顶尖和垫底位置被其他正在经历好运或霉运的公司所取代。这种上下波动是正常生命起伏的一部分，它并不意味着所有公司很快就会流于平庸。

让我们考虑一个具体的例子。假设某个行业里有三种公司：好公司，差公司，普通公司。每家好公司按照投资回报率计算的平均利润率为 40%。每家差公司的平均利润率为 20%。每家中等公司的平均利润率为 30%。每一种公司的数量相等。

每家公司的平均利润率是它的“能力值”。每家公司在任意一年的利润率等于其能力值+6%、+3%、-3% 或者-6%。因此，能力值为 40% 的公司在任意一年拥有 46%、43%、37% 和 34% 利润率的可能性是相等的。它们的年利润率在能力值附近随机波动。我并不是说现实中的能力永远不会发生变化。我只是在这里做出一种假设，以说明即使能力不变，回归也是有可能发生的。

这些假设用于生成两年的数据，我将这两年称为 1920 和 1930 年。在每一年，有 12 种具有同等可能性的利润率观测值：46%、43%、37%、36%、34%、33%、27%、26%、24%、23%、17% 和 14%。

根据西克里斯特的做法，假设我们根据这些公司的利润率观测值将其分成四个部分。最优的部分包含利润率观测值为 46%、43% 和 37% 的公司。次优的部分包含利润率为 36%、34% 和 33% 的公司。第三好的部分为 27%、26% 和 24%。垫底部分是 23%、17% 和 14%。

均值回归之所以发生，是因为利润率观测值远离均值的公司倾向于拥有更加接近均值的能力值。在这种情况下，他们在其他任意一年的利润率将更加接近均值。在我们的例子中，利润率观测值位于前 1/4 的公司拥有 46%、43% 和 37% 的利润率。它们的平均利润率是

42%。不过，所有这些公司都具有 40% 的能力值。它们排在所有公司前 1/4 的年份里拥有 42% 的平均利润率，但它们在其他任意一年的平均利润率期望都是 40%。表 9.1 显示了这些公司根据 1920 年利润率被分成四个部分时的收敛情况。

表 9.1：根据 1920 年利润率形成的四个部分的平均利润率

	1920	**1930**
第一部分，1920	42	40
第二部分，1920	34	33
第三部分，1920	26	27
第四部分，1920	18	20

1920 年位于前两个部分的公司倾向于在 1930 年拥有更加接近平均水平的利润率。1920 年位于后两个部分的公司同样倾向于在 1930 年拥有更加接近平均水平的利润率。

这种回归并不取决于我们使用哪一年将公司分成四个部分。不管使用哪一年分组，利润率都会在另一年出现回归。表 9.2 显示了根据 1930 年利润率分组时的收敛情况。

表 9.2　根据 1930 年利润率形成的四个部分的平均利润率

	1920	**1930**
最优部分，1930	40	42

次优部分，1930	33	34
第三好部分，1930	27	26
垫底部分，1930	20	18

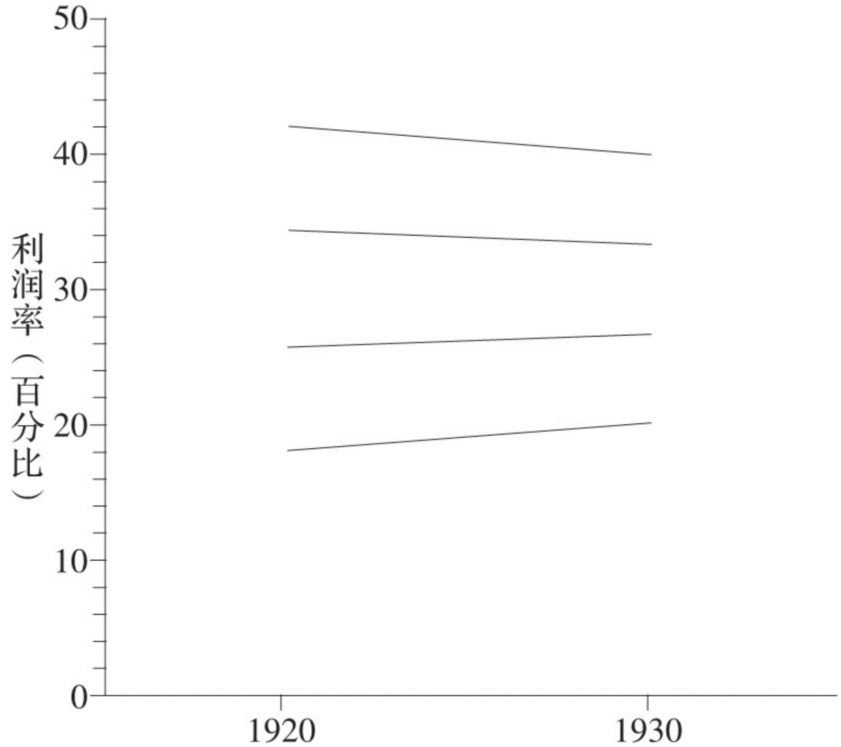


图 9.1　根据 1920 年利润率形成的四个部分在 1930 年出现回归

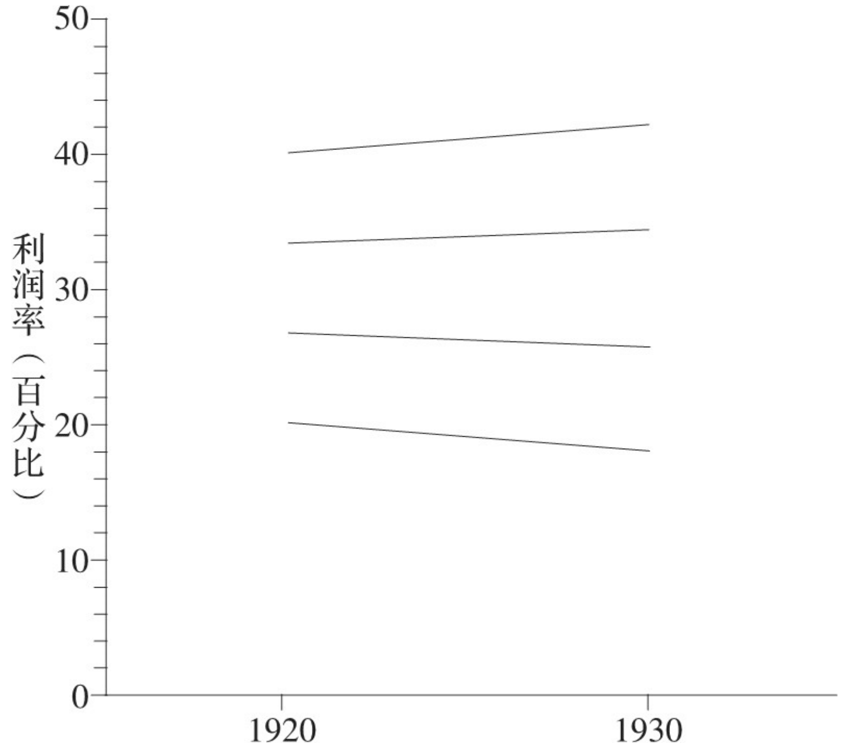


图 9.2　根据 1930 年利润率形成的四个部分在 1920 年出现回归

图 9.1 以图形方式说明了如果我们像西克里斯特那样使用 1920 年划分四个部分，利润率将在 1930 年出现回归。图 9.2 说明了如果我们使用 1930 年划分四个部分，利润率将在 1920 年出现回归。

这里面绝对不涉及能力值的收敛。我们假设每家公司每年的能力值是相同的。回归之所以发生，仅仅是因为利润率在能力值附近随机波动。

异常的父母通常拥有不太异常的孩子，反之亦然；同样的道理，不管我们沿着时间前进还是后退，利润率都会出现回归。观测到的回归只能证明当利润率在能力值附近波动时，观测到的利润率差异大于实际能力值的差异。正如霍特林所说，“这些图表实际上只能证明相关利润率存在波动趋势。”10 年的工作成果随着这句话流进了下水道。

古老的谬误何曾消亡?

虽然霍特林对西克里斯特的错误进行了明确的剖析，但是这个错误并没有消失。1970 年，一位著名的政治经济学家写道:

> 一部完全被人遗忘的、具有类似主题的早期经验性作品拥有一个意味深长的名字：《平庸在商业中的胜利》。这本书由贺拉斯·西克里斯特撰写，1933 年由西北大学经济研究局出版。这本书包含一个详细的统计学论证过程，证明了随着时间的推移，平均而言，最初表现出色的公司会出现退步迹象，而最初表现糟糕的公司则会出现改善。

遗憾的是，天真的作者并没有意识到西克里斯特的结论没有产生任何影响的原因。他只记得西克里斯特，忘记了霍特林。

20 世纪 80 年代，一位诺贝尔奖获得者在一本投资教材中指出，“最终，经济力量将导致不同公司利润率和增长率的收敛。”为了支持这种说法，他考察了 1966 年利润率最高和最低的公司。14 年后，1980 年，两个群体的利润率更加接近均值。他得意地总结道：“对总体均值的收敛是显而易见的……这种现象无疑是真实的。”多么似曾相识的说法！和 50 年前的西克里斯特一样，他并没有考虑到这种收敛仅仅来自均值回归的可能性。

几年后，另外两位杰出的金融教授（其中一位也是诺贝尔奖获得者）犯下了完全相同的错误。他们发现了数据中的收入回归现象。和西克里斯特一样，他们将其完全归因于竞争力量:

> 在竞争环境中，行业内和跨行业的利润率都会向均值靠拢。其他公司最终会模仿一家公司高于正常利润率所依靠的

> 创新产品和新技术。破产或被收购的前景将激励利润率较低的公司将资产分配到更有成效的地方。

竞争力量也许的确存在。不过，它们显然不是事情的全部。部分原因（甚至全部原因）在于“收入相对较高的公司更有可能正在经历好运，而不是霉运”这一单纯的统计性观察。

这种反复出现的错误的另一个版本出现在一本书及其书评之中，二者均出自著名经济学家的手笔。他们认为，所有国家的经济增长率随时间收敛。他们完全忽略了均值回归在这种收敛中起到的作用。米尔顿·弗里德曼（Milton Friedman）写了一篇恰当的评论，题目叫做《古老的谬误何曾消亡？》

> 这本书的作者和书评作者都是优秀的经济学家，对现代统计方法了如指掌，他们竟然没能认识到自己陷入了回归谬误，对此我感到很吃惊……不过，考虑到这种谬误在大众讨论和学术研究中的普遍存在，我也许不应该感到吃惊。

既然著名统计学家、金融教授和经济学家都能忽略均值回归，其他人又怎么能幸免呢?

道指落榜生

道琼斯工业平均指数（“道指”）是代表美国最优秀公司的 30 只蓝筹股票的平均价格。根据道琼斯公司的说法，这些“重要公司以其产品或服务的质量和被广泛承认著称，拥有强劲而成功的增长历史”。

一个平均委员会定期改变道指中的股票。有时，这是因为公司与另一家公司合并或者被另一家公司收购后股票停止了交易。有时，公司进入了困难时期，不再被视为真正的蓝筹股票。这些"落榜"公司被更加成功的公司所取代。

例如，1991 年 11 月 1 日，家得宝取代了西尔斯。西尔斯是一个具有传奇色彩的美国成功故事，它从一家销售手表、玩具、汽车和易组装房屋等各种商品的目录邮购公司发展成了美国最大的零售商。西尔斯在道指中停留了 75 年，但它目前在与沃尔玛、塔吉特和家得宝等折扣零售商的竞争中陷入了困境。西尔斯的收入和利润开始下降，股价也在过去的 6 个月里下跌了将近 50%。家得宝则在住宅建设和装修的带动下急速发展，每过 56 个小时就要开一家新店。该公司的工具与西尔斯具有传奇色彩的克拉夫茨曼工具存在直接竞争关系，而且它似乎正在赢得这场销售战争。家得宝的股票价格过去 6 个月上涨了 50%。

当一家正在衰落的公司被一家蒸蒸日上的公司取代时，你认为哪只股票会在接下来表现得更好——进入道指的股票，还是离开道指的股票？如果你能考虑到均值回归，你就会知道，被踢出道指的股票未来的表现很可能会优于取代它的股票。

这是违反直觉的，因为我们很容易将伟大的公司与伟大的股票混淆在一起。假设你发现了一个历史上长期具有丰厚稳定利润的伟大公司（让我们称之为利恩米恩）。利恩米恩是一个优秀的投资项目吗？答案取决于股票价格。每股 10 美元的投资有吸引力吗？100 美元呢？1000 美元呢？对于某些价格，股票太贵了。对于某些价格，股票太便宜了。不管公司多么优秀，我们都需要首先了解股票价格，然后再去判断它是不是一个具有吸引力的投资项目。

遇到困难的公司也是一样的道理。假设一家不幸叫做"涤纶西装"的公司处于下行死亡螺旋之中。公司目前每股分红 1 美元，这个数字预计每年稳步下降 5%。谁会购买如此失败的股票呢？如果价格合适，谁会不买呢？你会用 5 美元购买每年的股息分别为 1 美元、95 美

分、90 美分……的股票吗？如果 5 美元的价格无法打动你，1 美元呢？10 美分呢？

让我们回到道指添加和删除的股票上来。投资者面对的问题不是进入道指的公司目前是否比它们所取代的公司更加成功，而是哪些股票是更好的投资品。进出道指的股票都是耳熟能详的公司，都会受到数千名投资者的密切关注。1999 年，投资者非常清楚，家得宝表现出色，西尔斯表现糟糕。这种观念一定会体现在它们的股价上。这就是家得宝股价上涨 50%、西尔斯股价下跌 50% 的原因。

不过，由于均值回归，被移出道指的公司通常不像它们最近的表现所暗示的那样可怕，取代它们的公司通常也不像它们表现得那样优秀。在这种情况下，落榜的股票通常拥有不合理的低价，入榜的股票通常拥有名不副实的高价。当一家表现糟糕的公司向平均水平回归时，其股价会上涨；当一家表现出色的公司向平均水平回归时，其股价会下跌。这种观点意味着同添加到道指中的股票相比，被道指删除的股票通常会表现得更好。

在被踢出道指五年半以后，西尔斯在 2005 年被凯马特收购。如果你在西尔斯被道指删除以后立即购买它的股票，那么到西尔斯被凯马特收购的时候，你的总回报率将达到 103%。在同样的五年半时间里，取代西尔斯的家得宝下跌了 22%。标准普尔 500 股价指数这段时期的回报率是-14%。西尔斯离开道指后的回报率高于平均水平，而家得宝进入道指后的回报率低于平均水平。（凯马特和西尔斯的合并效果非常糟糕，但那又是另一个故事了。）

西尔斯与家得宝的对比是一起孤立事件，还是“道指落榜生表现优于道指入榜生”这一系统性模式的一部分呢？实际上，1999 年发生了四次替换：家得宝、微软、英特尔和西南贝尔公司取代了西尔斯、固特异轮胎、联合碳化物和雪佛龙。家得宝、微软、英特尔和西南贝尔公司都是伟大的公司，但这四只股票在接下来的 10 年里表现都很糟糕。

假设你在 1999 年 11 月 1 日四次替换发生那天向道指添加的四只股票各投资 2500 美元。共计 1 万美元。这是你的入榜投资组合。你还建立了一个落榜投资组合，向道指删除的四只股票各投资 2500 美元。

表 9.3 显示了替换发生后的 10 年时间里这些投资组合的表现与标普 500 的对比。10 年后，标普 500 下跌了 23%。入榜投资组合表现更加糟糕，下跌了 34%。相比之下，落榜投资组合上涨了 64%。

表 9.3　1999 年 11 月 1 日添加和删除的股票

	初始投资组合	五年后	十年后
入榜投资组合	10000 美元	6633 美元	6604 美元
落榜投资组合	10000 美元	9641 美元	16367 美元
标普 500	10000 美元	8295 美元	7652 美元

这仍然只是 1999 年道指的四次替换而已。也许 1999 年是不同寻常的一年，其他年份的替换具有不同的表现？非也。2006 年的一项研究考察了 1928 年 10 月 1 日道琼斯 30 只股票平均指数诞生以来的所有 50 次更改，发现在 32 次更改中，被删除股票的表现优于替代它们的股票，在 18 次更改中，被删除股票的表现不如替代它们的股票。被删除股票的组合每年比新增股票的组合好 4%。经过 78 年的积累，这是一种巨大的差距。100 美元新增股票的投资组合将在 2006 年增长到 16 万美元；100 美元被删除股票的投资组合将增长到 330 万美元。

由于表现糟糕而被踢出道指的公司比取代它们的好公司更加适合投资。

冠军室息

大多数体育迷都相信“冠军窒息”现象——取得某种优异成绩的运动员随后的表现往往令人失望。显然，人们会付出不同寻常的努力，以取得不同寻常的成绩。不过，当他们登顶时，他们对失败的担忧会使他们所担忧的失败变成现实。同样的现象似乎也存在于许多职业领域。

在罗杰·马里斯（Roger Maris）1961 年以 61 个全垒打超越贝比·鲁斯以后，他在 1962 年和 1963 年分别只打出了 33 个和 23 个全垒打。在马丁·斯科塞斯（Martin Scorsese）1976 年制作了《出租车司机》以后，他在 1977 年制作了糟糕透顶的《纽约，纽约》。

最有名的例子是《体育画报》封面诅咒。在俄克拉荷马连续赢得 47 场大学橄榄球比赛胜利之后，《体育画报》刊登了《俄克拉荷马为何战无不胜》的封面故事。俄克拉荷马在下一场比赛中以 21:28 输给了圣母大学。经过这次溃败，人们开始注意到，出现在《体育画报》封面上的运动员明显受到了诅咒，无法在之后的比赛中取得同样优秀的表现。2002 年，《体育画报》刊登了关于诅咒的封面故事，使用了黑猫的图片和可爱的文字说明“没人愿意登上的封面”。最近，我们又有了“麦登诅咒”：出现在视频游戏《麦登橄榄球》封面上的橄榄球选手无法取得像之前那样优秀的表现。

不好意思，让你们这些真正相信诅咒的人失望了。实际上，这仅仅是均值回归而已。

运动员的表现是对技能的不完美测量，因此它会向均值回归。表 9.4 显示了 2010 年平均击球率最高的十位职业棒球大联盟选手。乔希·汉密尔顿（Josh Hamilton）的平均击球率最高，为 0.359。你认为乔希是一个在 2010 年赛季表现令人失望的、击球率为 0.400 的击球手呢，还是在 2010 年表现出色的、击球率为 0.300 的击球手呢？2010 年排在前十名的击球手都在这一年取得了超出职业生涯平均水平的表现。总体而言，这十位选手职业生涯的平均击球率约为 0.301，他们在 2010 年的平均击球率则是 0.327。

在 20 项比较中，这些选手只有 3 次在 2009 年或 2011 年的表现优于 2010 年（用斜体显示）。在 20 项比较中，前十位击球手 2010 年的平均击球率在 2009 年和 2011 年有 17 次出现了向均值的回归（回归是一种趋势，不是必然的）。

回归并没有描述“能力”随时间的变化，比如公司被竞争削弱或者棒球选手的状态随着年龄的增长而下滑。回归是由“表现”相对于“能力”的波动导致的，因此远离均值的表现反映了更加接近均值的能力。所以，不管我们向前查看下一年，还是回头查看上一年，我们都会观察到棒球选手的回归现象。

表 9.4　2010 年平均击球率最高的 10 位选手 2009 年和 2011 年的表现

	2009	**2010**	**2011**	**职业生涯**
乔希·汉密尔顿	0.268	0.359	0.298	0.308
卡洛斯·冈萨雷斯	0.284	0.336	0.295	0.298
米格尔·卡布雷拉	0.324	0.328	0.344	0.317
乔·莫尔	0.365	0.327	0.287	0.323
乔伊·沃托	0.322	0.324	0.309	0.313
艾德里安·贝尔特	0.265	0.321	0.296	0.276
奥马尔·因方特	0.305	0.321	0.276	0.275
罗宾逊·卡诺	0.320	0.319	0.302	0.308
比利·巴特勒	0.301	0.318	0.291	0.297

特罗伊·托洛维斯基	0.297	0.315	0.302	0.293
平均	0.305	0.327	0.300	0.301

根据“平庸胜利”谬误，我们会得出优秀选手和队伍技能退步的结论。实际上，正确结论是，任何给定赛季表现最优秀的人通常不像他们耀眼的记录显示的那样拥有出色的技能。大多数人的运气比较好，这使他们在某个赛季的表现优于上一个赛季和下一个赛季——而在上一个赛季和下一个赛季，他们的顶尖位置将被其他人取代。

《体育画报》诅咒和约翰·麦登诅咒是均值回归的极端案例。当一名选手或一个团队做出足以登上《体育画报》或《麦登橄榄球》封面的优异表现时，他们接下来的表现几乎一定会出现退步。考虑到运气在运动员成功的表现中一定会起到的作用，凌驾于其他人之上的选手或团队几乎一定受到了运气的帮助——身体健康、幸运的反弹球、可疑的裁判。好运不可能永远持续下去，非凡的成功也是如此。

寻找学院院长和灵魂伴侣

几年前，一所小型私立大学在全国范围内寻找一位新的学院院长。学校认为所有内部候选人都不够优秀。遴选委员会根据数百名外部申请人的简历和介绍信选出了几个人，邀请他们在中立机场参加面试，以免泄密。令遴选委员会印象最为深刻的三位候选人被邀请到校园里，在两天时间里与教员、行政人员、职员和学生见面。

遴选委员会对这三位优秀候选人极为热情。不过，同候选人到来之前的大肆宣传相比，每位候选人的实际表现都很令人失望。人们谣传说，遴选委员会已经确定了人选，他们故意请来了两位失败者，以

表明他们中意的人是最优秀的候选人。不过，关于谁是委员会中意的人，人们产生了极大的分歧。

你觉得这个例子存在均值回归现象吗？没有人能够仅仅根据简历、介绍信和机场面试知道每位候选人的优秀程度。你认为看上去最优秀的三位候选人实际上比他们的表现更加优秀，还是不像他们表现得那样耀眼？如果一个人比他表现得还要好，但他仍然能够在候选人之中排在前三位，这样的事情实在是太少见了。失望几乎是无法避免的，因为表现最为优秀的三位候选人几乎一定不像他们看上去那样出色。均值回归现象还可以解释内部候选人为何处于固有的劣势地位。一个在大学里工作了二三十年的人并没有太多可以隐藏的优缺点。和基本处于未知状态的外部候选人不同，对于内部候选人，你对他的了解和他未来的表现基本是一致的。

有一年，我在我的统计课堂上讨论了这段经历。课后，一名学生找到我，向我讲述了一个不同寻常的巧合。前一天晚上，他和他的父亲通了电话。他的父亲是另一所大学的社会学教授，他抱怨说，当他们邀请最优秀的教员岗位申请人来到校园时，这些申请人通常不像纸面上和之前的短暂面试中表现得那样令人激动。这名学生说，他会在下午给父亲打电话，简单谈谈均值回归现象。

我们对灵魂伴侣的寻找也是同样的道理。每个人都在寻找不同的东西，我们可以称之为“魅力”。在我们见到的几十人、几百人或者几千人之中，少数人会脱颖而出。当我们看到魅力的迹象并鼓起勇气深入了解这个人时，我们通常会失望。坏消息是，这是正常的。好消息是，我们需要继续寻找。值得深思的消息是，对方很可能和我们有着同样的感受。

实际上，在生活中，我们几乎每天都会遇到均值回归现象。我们应当努力预测和识别这种现象，做到不被这种现象欺骗。不要忘记西克里斯特的故事。

×如何轻松识破一本正经地胡说八道

当学术能力或运动能力等特点得到不完美测量时，观测到的表现差异会夸大实际能力差异。表现最优秀的人与平均水平的距离很可能不像看上去那样遥远，表现最为糟糕的人也是如此。因此，他们随后的表现将会朝着均值回归。

这并不意味着表现最优秀的人受到了诅咒。实际上，这只是因为他们的出色表现得到了好运的帮助。均值回归也不意味着能力向均值收敛、大家很快就会具有平均水平，它仅仅意味着极端表现在经历好运和霉运的群体之间轮换。均值回归也不意味着成功和不成功的公司将收敛到令人沮丧的平庸状态。

第 10 章
如何转变运气?

在《宋飞正传》的“反面”一集中，乔治承认，他的生活之所以与他的设想完全相反，是因为他所做的每个决定都是错误的。杰里告诉乔治，解决方法是将他正常想做的事情反过来。因此，当乔治看到一个美女独自吃饭时，他应该走过去说：“我叫乔治，无业，目前和我的父母生活在一起。”令人难以置信的是，这种做法成功了。当他应聘纽约扬基队的工作时，他指责老板乔治·施泰因布伦纳（George Steinbrenner）毁掉了这支队伍。他当场受到了聘用。

与此同时，伊莱恩的生活变得一塌糊涂，她失去了男友、工作和公寓。在这一集的结尾，伊莱恩抱怨说，她已经变成了乔治。相比之下，杰里的生活和之前没什么两样。克雷默给杰里起了“正负相抵”的外号，因为杰里遇到的每件坏事都会被一件好事抵销。当一场独角戏演出被取消时，杰里立即在同一个周末获得了同样价钱的工作。当他打扑克时，他的输赢局数相互抵销，因此不赔不赚。当他的女友抛弃他时，杰里镇定自若，因为他相信新的女友很快就会出现。

杰里不是唯一感觉事情会正负相抵的人。许多人相信，每当硬币正面朝上时，背面朝上的可能性就会增长，因为长期来看，正面和背面出现的可能性一定是相等的。同样的道理，棒球经理会用替补击球手换下连续四次实现安打的选手，因为这名选手已经到了该出局的时候。

这种观点的一种变化形式是认为世界上好事和坏事的供应量是固定的。我曾经看到一名射手在赛季初期的大学橄榄球比赛中葬送了三个三分球和一个加分球。电视评论员说，教练应当对这些失误感到高兴，因为接下来的几个星期会有几场艰难的比赛。评论员说，每位射手在赛季里都会出现一些失误，在年初将这些失误“用掉”是令人欣喜的。

一位读者给专栏作家玛莉莲·沃斯·莎凡写信，说他参加了许多工作面试，但是没有收到任何录用通知。他希望录用的可能性能够随着拒绝次数的增加而增加。另一位读者向玛莉莲写道：

> 在我参加的一场关于消防安全的讲座上，主讲人说：“我知道你们中的一些人会说，你们在家里生活了25年，从未经历任何类型的火灾。对此，我想说的是，你们过去比较幸运。不过，这意味着你们和下一场火灾的距离不是越来越远，而是越来越近。”

西弗吉尼亚州最高法院首席法官曾经开车前往南达科他州参加一场司法会议。他解释道：“在我的一生中，我坐了许多次飞机。我已经用掉了我的统计英里数。只要还有其他可行的替代方案，我是不会坐飞机的。”

这种常见的推理过程基于错误的平均定律。一些人认为，如果抛1000次硬币，一定会有500个正面和500个背面；因此，如果前十次、前五十次或者前一百次出现的正面多于背面，那么为了使结果得到平衡，接下来出现的背面一定多于正面。这是一种普遍存在的观点，但它是错误的。硬币无法控制自己的落地方式。如果硬币是完好的，并且得到公平抛掷，那么不管上一次或者之前999次抛掷的结果如何，正面和背面接下来出现的可能性都是相等的。

不过，许多赌徒仍然相信错误的平均定律，因为他们急于在混乱的随机结果中找到一种盈利模式。当轮盘连续多次转出黑色数字时，

总会有人急切地在红色上下注，认为平均定律将会发挥作用。其他人则匆忙地在黑色上下注，试图赶上这一波黑色运势。庄家愉快地接受双方的赌注，他们知道，未来的旋转并不取决于过去的结果。

最具戏剧性的轮盘赌之一发生在 1913 年 8 月 18 日的蒙特卡洛。当时，一张赌桌上反复出现黑色结果。当黑色连续出现 10 次时，赌桌围满了投注红色的人，他们相信平均定律将会为他们带来回报。不过，黑色仍然不断地出现。在连续出现 15 次黑色以后，人们近乎疯狂地挤到赌桌前，以便将更多赌注压到红色上。接下来出现的仍然是黑色。当黑色第 20 次出现时，绝望的赌徒将他们剩下的所有筹码压在红色上，希望挽回一部分损失。当这轮离奇的轮盘赌结束时，黑色连续出现了 26 次，赌场收获了数百万法郎。

一定会有一些失败者认为这场赌局受到了操纵。也许事实的确如此。不过，在蒙特卡洛和其他地区的公平赌局中，基于平均定律的投注策略是无效的。实际上，有效的少数投注体系基于相反的原则——设备的物理缺陷导致一些数字出现的频率高于其他数字。19 世纪末，英国工程师威廉·贾格尔斯（William Jaggers）雇用了六名助手，用一个月的时间记录了蒙特卡洛轮盘赌的获胜数字。他不是根据事物的平均定律避开这些数字，而是投注这些数字，因为他觉得轮盘的不完美将使这些数字的出现频繁继续高于其他数字。他赚了将近 12.5 万美元——相当于今天的 600 多万美元。后来，赌场发现了问题，开始在夜间调整轮盘，贾格尔斯的幸运之旅这才结束。

人类与硬币和轮盘不同，他们拥有记忆，而且关心胜负问题。不过，棒球选手的安打概率并不会因为上一次的失误而得到提升。连续四次出局也许仅仅是因为运气不好，导致直球直接砸进了外野手的手套。这种糟糕的运气并不能确保下次击球出现好运。如果不是运气的原因，那么选手的糟糕表现可能来自身体问题。不管是哪一种情况，连续四次出局的棒球选手不一定在下次击出安打，连续四次安打的选手也不一定在下次出局。如果说我们能够得到什么结论的话，连续四次安打的击球手很可能比连续四次出局的击球手更加优秀。

类似地，三分球失误不一定会得到成功的补偿。糟糕的表现可能仅仅意味着射手不是很优秀。面试被拒并不会提高被录用的可能性。相反，这只能更加证明这个人不合格或者在面试中表现糟糕。没有发生火灾并不会提高发生火灾的可能性；它可能仅仅说明房主比较谨慎，不会把纸张和布料放在火炉附近，不会把金属放进微波炉，不会在离家时开着火炉，不会在吸烟时睡着。每一次安全的飞机旅行并不会提高下次飞机坠毁的可能性。

霉运不会提高好运的可能性，反之亦然。每一次失败不会提高成功的可能性，反之亦然。它可能仅仅是随机性的一种表现而已。

×如何轻松识破一本正经地胡说八道

当我们经历糟糕的运气时，我们希望自己能够转运。我们的霉运不可能永远持续，但发生在我们身上的坏事并不会自动提高发生好事的可能性。要想改变运气，我们通常需要改变自己的行为。例如，如果我们在找工作时不断遭到拒绝，我们应当考虑如何更好地表现自己，或者考虑申请不同的工作。

正负相抵只是一个笑话，不是值得信赖的规律。

第 11 章
德克萨斯神枪手

奇斯威克是西伦敦一个美丽富饶的郊区，其境内的泰晤士河是著名赛艇比赛的一部分赛道。牛津和剑桥之间具有传奇色彩的年度赛艇比赛将终点设在奇斯威克桥附近。奇斯威克还培养出了小说家 E. M.福斯特（E. M. Forster）、毕沙罗画家家族、摇滚乐手皮特·汤森（Pete Townshend）、女演员海伦·米伦（Helen Mirren）以及伦敦最大、最古老的啤酒厂——如今以“伦敦之巅”闻名的富勒啤酒厂。

第二次世界大战期间，这个宁静的田园小镇遭到了多枚德国炸弹的袭击，包括 1944 年 9 月 8 日星期五晚 7 点之前毫无预兆的一枚炸弹。斯泰夫利路学校 64 岁的管理员罗伯特·斯塔布斯（Robert Stubbs）一边穿越学校的运动场，一边思考一些杂务以及即将到来的降雨。突然，他被抛到了 6 米以外的草地上。当他站起来的时候，他发现被震飞的不仅仅是他的身体。到处都是残骸和碎石。路旁的许多树木被击倒，甚至被炸成碎片。斯泰夫利路的房屋有着 22 厘米厚的墙壁，但是仍然有 11 幢房屋被毁，还有 27 幢房屋遭到严重破坏，无法居住。一名士兵死在了拜访女友的路上。一个三岁的孩子死在婴儿床上，因为爆炸吸走了她肺部的空气。一名妇女在倒塌的房屋中遇难。还有 17 人在倒塌的房屋中严重受伤。要不是许多人在那个夏天由于担心受到此类袭击而撤离奇斯威克，伤亡人数还会进一步增长。

一位目击者表示，当时的声音听上去就像煤气总管爆炸一样。一位政府官员也表示，这完全有可能是煤气总管爆炸。这种说法成了官方的解释。奇怪的是，奇斯威克爆炸发生 15 秒以后，伦敦另一个地区发生了另一起“煤气总管爆炸”。市民产生了怀疑，并用嘲讽的口气将类似爆炸称为“飞行的煤气管道”。他们的怀疑是有道理的。英国政府最终承认，奇斯威克遭到了从德占荷兰发射的 V-2 火箭的袭击。

V-2 火箭重 8000 磅，飞行速度为每秒 900 米，对地面的冲击相当于 50 辆以每秒 96 公里的速度前进的、重达 100 吨的火车头。落地以后，火箭的弹头发生爆炸，在斯泰夫利路形成了一个 9 米宽、3 米深的弹坑。冲击波为 400 米范围内的一切事物带来了沉重的打击。

在二战最后阶段，德国军队向伦敦发射了数千枚 V-1 和 V-2。V-1 是 7.5 米长的无人飞机，用内部陀螺仪控制飞行路径。这些飞机以每小时 480 公里的速度前进，同时发出嘎嘎声。从德国占领的法国发射场地到伦敦，它们需要飞行大约 30 分钟。飞机搭载了一个小型风车装置。当飞机向伦敦飞行时，这个装置可以根据风车的转动次数估计 V-1 前进的距离。经过预先设置的转动次数以后，V-1 将会朝着地面大角度俯冲下去。

V-2 是最早的远程战斗火箭和最早抵达外层空间的火箭。这种火箭以每小时超过 4800 公里的速度飞行，几分钟就能抵达伦敦。由于 V-2 的飞行速度是音速的 4 倍，而且飞行高度超过 80 公里，因此它们几乎无法用战斗机或高射炮拦截。每枚火箭携带一吨弹头，能够毁掉 400 米范围内的一切事物。

从 1944 年 6 月到 1945 年 3 月，伦敦遭到了 2419 枚 V-1 和 517 枚 V-2 的袭击。V-1 在伦敦上空飞行时会发出“嘟嘟嘟”的单调声音，可以提醒人们炸弹来了。而且，最初的 V-1 存在设计缺陷，向地面俯冲时会关掉发动机。伦敦人知道，当“嘟嘟嘟”的声音停止时，他们就要寻找藏身之处了。同 V-1 相比，V-2 更加可怕，因为它的飞行速度高于音速，而且不会在袭击之前发出提醒。更加怪异的是，在火箭袭击地面以后，人们才会听到火箭俯冲时发出的呼啸之声。呼啸声结束后，

人们还会听到上层大气的音爆，仿佛时光出现了逆转。先是巨大的爆炸声，然后是来自天空的、震耳欲聋的呼啸声，然后是呼啸声的起始地点传来的音爆。

在人们星期日做礼拜的时候，一枚 V-1 击中了威灵顿兵营的卫兵教堂。这个教堂是一座雄伟的建筑，拥有加固的屋顶。不过，炸弹仍然炸飞了屋顶，将巨大的建筑劈成两半，导致 119 人丧生，141 人受伤。还有一枚 V-2 击中了伦敦一家沃尔沃斯商店。整个建筑被彻底摧毁，变成了商店地下室里的一堆碎石。168 名购物者丧生，其中一些人死于爆炸的冲击波，一些人死于建筑的倒塌。当炸弹降落在居民区时，整个街区都可能被摧毁。

所有袭击导致数万人死伤，数百万住宅、办公室、工厂被毁。平民精神状态受到的影响也许更加严重。不管是去教堂，还是去购物，或是待在家里看电视，哪里都是不安全的。超过一百万人离开伦敦。

虽然 V-1 和 V-2 造成了巨大的破坏，但它们的准确率非常有限。例如，V-1 的风车系统很容易被风干扰，因为风会使风车加速或者减速。德国人将 V-1 对准伦敦塔桥，但是只要 V-1 降落在伦敦范围内，他们就会感到满意。实际上，数百枚飞弹落在了田野和海洋里，没有造成任何破坏。

不过，许多伦敦人相信，德国人对准了某些街区。一些区域似乎存在炸弹聚集现象，另一些区域则很少被炸弹袭击。这意味着德国人瞄准了一些街区，避开了另一些街区。在寻找这些数据聚集现象的原因时，一些人提出，德国间谍可能生活在德国人没有轰炸的区域。他们轻率地认为，如果存在某种模式，那么它一定拥有某种原因。实际上，这是不一定的。

即使是随机数据，也会出现聚集现象。我将一枚 25 美分硬币抛了 10 次，下面是我得到的结果：

背面 正面 正面 背面 背面 正面 正面 正面 正面 正面

在这个小实验的后半段，5 个正面聚集在了一起。发生了什么事情？我是否改变了我所面对的方向，或者坐在了另一把椅子上？不。没有发生任何特别或不同寻常的事情。当我们看到数据聚集时，我们自然而然地认为发生了某种特别的事情——这些正面（或背面）成串出现是有原因的。实际上，这是没有原因的。即使是完全随机的抛硬币实验，也会出现数据的聚集。如果你不相信，你可以抛十次硬币。即使每次抛掷是完全随机的，你也有 47% 的概率得到至少四个连续正面或连续背面。

随机投掷的炸弹落点也是同样的道理。一些区域会出现炸弹的聚集，另一些区域不会出现炸弹，这是无法避免的事情。英国统计学家 R. D.克拉克（R. D. Clarke）对南伦敦一个区域的数据进行了分析，这个区域受到了 537 枚炸弹的袭击。他将这个区域分成 576 个一平方公里的正方形，然后统计落在每个正方形里的炸弹数量。接着，他将炸弹分布与 537 枚炸弹随机降落在南伦敦时的分布期望进行了比较。表 11.1 显示了他的发现。

表 11.1　随机降落的炸弹

正方形里的炸弹数量	**正方形的预期数量**	**正方形的实际数量**
0	227	229
1	211	211
2	98	93
3	31	35
4	7	7
5 及以上	2	1
总计	576	576

如果炸弹是随机降落的，预计将会有227个正方形不会受到炸弹袭击。实际上，229个正方形躲过了攻击。类似地，我们预计211个正方形拥有一枚炸弹；实际上，221个正方形拥有一枚炸弹。总体而言，观测到的分布与随机轰炸的预期结果非常类似。在随机轰炸的情况下，实际差异和观测到的差异一样大甚至更大的可能性是95%。这些观测到的炸弹聚集绝对没有任何意义。它们与炸弹随机降落时的预期相符。

即使是完全随机的数据也会出现数据聚集现象，因此我们没有必要根据这种现象徒劳地寻找异想天开的解释。遗憾的是，人们很难抗拒“每一种模式一定有原因”这一想法的诱惑。

癌症聚集恐慌

20世纪70年代，失业的流行病学家南希·韦特海默（Nancy Wertheimer）开车在丹佛市穿行，寻找19岁之前死于癌症的人生活过的家庭。她在寻找这些家庭的某种共同点——尽管她并不知道这种共同点是什么。

她发现，许多癌症受害者的家庭靠近大型输电线。她的生活伴侣、物理学家埃德·利珀（Ed Leeper）带着她开车在丹佛市穿行，测量每个家庭与不同类型的电线和变压器之间的距离，并对电磁场的强度做出假设，以估计电磁场（EMF）的辐射量。他们对没有儿童癌症受害者的丹佛家庭样本进行了同样的估计。他们承认，他们并不了解输电线可能导致癌症的原因，但他们仍然总结说，生活在大型输电线附近的孩子患癌的风险将提高一到两倍。

他们发表的报告以及一些可怕的故事引发了全国范围的恐慌。一个自杀的人曾经生活在输电线附近。一些不再下蛋的母鸡生活在输电

线附近。一个名叫保罗·布罗德（Paul Brodeur）的新闻工作者在《纽约客》上撰写了三篇文章，指出康涅狄格州吉尔福德市一个发电厂附近的居民之中出现了异常众多的癌症病例，加州弗雷斯诺市输电线附近一所小学的员工也出现了 15 个癌症病例。布罗德将这些文章做了些许修改，放在了一本书中，书名是《输电线大揭秘：公用事业公司和政府如何试图隐瞒电磁场带来的癌症风险》。他提出了不祥的警告："由于暴露在输电线磁场之中，数千名毫不知情的儿童和成人将受到癌症的袭击，其中许多人将毫无必要地夭折。"

随之而来的全民疯狂为咨询师、研究员、律师和小型设备制造商带来了利润丰厚的业务，包括制造允许人们在家里测量电磁场的高斯计（电磁场读数较高的房间将被仅仅用于存储物资）。

和前面一样，问题在于，即使是随机数据，也可能存在聚集现象。即使癌症在人群中的分布是随机的，癌症受害者也存在地理聚集的可能性。为了说明这一点，我设计了一座虚拟城市，城市里的一万名居民生活在间隔均匀的住宅里，每个人有 1% 的患癌可能性。（我忽略了人们生活在家庭里以及癌症与年龄有关的事实。）我通过计算机模拟抛硬币的方式决定癌症受害者在这座虚拟城市里的分布。图 11.1 显示了我所得到的癌症地图。白色区域没有癌症受害者。

地图下方显然存在一个癌症聚集区。如果这是真实的城市，那么当我们开车在这些人居住的街区行驶时，我们一定能够找到某种特别的东西。也许，城市的少年棒球联合会球场坐落在附近。现在，如果我们将这些球场附近的居民患癌率与远离球场的居民患癌率进行比较，你觉得会有什么结果？球场附近将拥有更高的患癌率，这说明生活在少年棒球联合会球场附近会导致癌症。拆掉这些球场！

图 11.1 还显示了一个防癌堡垒，即无人患癌的一个城市区域。如果我们在这个没有癌症的街区行驶，我们一定会找到某种不同寻常的事物。也许，城市的水塔坐落在附近。现在，如果我们将水塔附近的居民患癌率与远离水塔的居民患癌率进行比较，水塔附近的患癌率当

然会更低一些。这就是我们一开始在这片街区进行寻找的原因——这里没有人患癌。

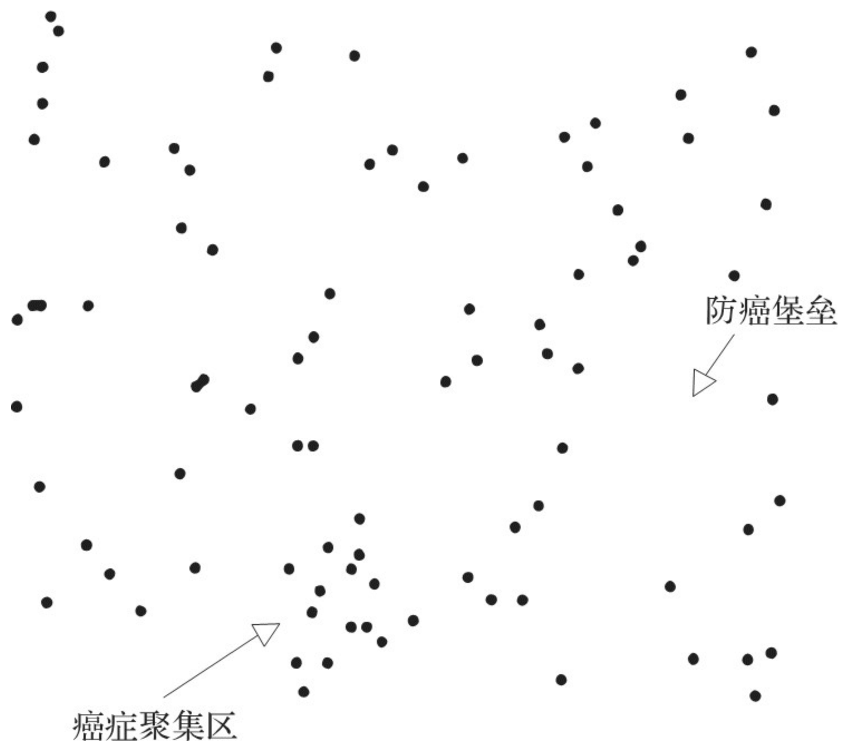


图 11.1　癌症地图

我们对少年棒球联合会球场附近和水塔附近的推理存在同样的问题。如果我们根据数据编造理论（少年棒球联合会球场导致癌症，水塔预防癌症），那么这些数据当然会支持这种理论！它怎么会不支持呢？我们会编造出一种与数据不符的理论吗？当然不会。根据创建理论时使用的数据来检验这种理论的做法是不公平的。我们需要新的数据。在新数据面前，这种理论也应当成立。

有时，数据聚集的确非常重要。1976 年夏天，在由一群退伍军人组成的宾州美国军团的年度会议上，一种神秘肺病袭击了 182 人，并导致 29 人死亡。经过大量侦察工作，科学家发现了导致这些人患上肺炎（现在叫做军团病）的，由空气传播的细菌。这些科学家还发现了比疾病最初爆发时使用的抗生素更加有效的抗生素。

对于军团病，人们根据数据聚集得到一种理论，并用新数据对其进行检验。如果没有理论和新数据，我们所拥有的仅仅是数据聚集而已。韦特海默和利珀发现的癌症受害者聚集是“德克萨斯神枪手”谬误的一种形式。在最初的版本中，一个没有射击技能的人用一把枪向谷仓的一面墙射出大量子弹，然后在弹孔最多的位置画上靶心。即使癌症的地理分布是完全随机的，也会出现巧合性的聚集——就像 100 次随机射击的弹孔会出现随机聚集一样。

要想进行有效的统计检验，研究人员应当首先画出靶子，然后发射子弹——首先论证输电线可能导致癌症的原因，然后比较有输电线和没有输电线的街区发生癌症的频率。

韦特海默和利珀这项研究的另一个问题是，他们对电磁场辐射量的估计可能会受到他们对研究结果预期的影响。有偏估计很容易将非常小的风险率提高一两倍。要想进行科学的研究，应当让中立的观测者在不知道哪些住宅有癌症受害者、哪些住宅没有癌症受害者的情况下对电磁场的辐射量进行估计。

在韦特海默和利珀的报告发表以后，玛丽亚·费克廷（Maria Feychting）和安德斯·阿尔鲍姆（Anders Ahlbom）开展了一项更加严格的输电线电磁场研究，他们找出了生活在高压输电线 300 米范围内的瑞典家庭。他们不是估测电磁场辐射量，而是使用详细的瑞典公用设施记录计算这些人 25 年来每年接受的电磁场辐射。经过对数据的细致分析，费克廷和阿尔鲍姆的结论是，接受辐射量最大的孩子患上儿童白血病的可能性是没有受到辐射的孩子的 4 倍。

费克廷和阿尔鲍姆的研究体现了另一种形式的德克萨斯神枪手谬误。冒牌神枪手不是先射击，再在最密集的区域画上靶心，而是向许多目标射出许多子弹。然后，他找到自己所击中的一个目标，用油灰盖住其他弹孔，然后刷掉一切痕迹，只留下他所击中的目标。

费克廷和阿尔鲍姆的研究考虑了多种癌症、多个年龄群体、多种电磁场衡量标准以及划分“无辐射”“轻度辐射”“中度辐射”和“重度辐射”的多个阈值，从而创造出了数百个目标。他们计算了将近 800 种风

险率，然后选择了儿童白血病。如果输电线完全是良性的，那么一些风险率将仅仅由于运气不佳而高于其他风险率。其他一些风险率将远低于平均水平。费克廷和阿尔鲍姆自然会选择报告最高的风险率，不去提及其他风险率。他们掩盖了没有命中的目标。

这种神枪手谬误形式也叫费曼陷阱，是以诺贝尔奖获得者理查德·费曼（Richard Feynman）的名字命名的。我们将在 13 章再次提到他。费曼请加州理工学院的学生计算他在走出教室时在停车场里看到的第一辆车拥有某个特定牌照（比如 8NSR261）的概率。加州理工学院的学生非常聪明，他们假设牌照中的每个数字和字母是相互独立的，然后迅速计算出了 8NSR261 的概率。他们的答案是 1.76 亿分之一。当他们得到答案时，费曼说，正确的概率是 1，因为他刚才在路上看到了这个牌照。如果某件可能性极低的事情已经发生，那么它就不再是小概率事件了。

不过，研究人员一直在做同样的事情。在考察数据之后预测数据的形式是很容易的，也是没有意义的。在我们的输电线例子中，在观测儿童白血病以后预测儿童白血病的做法是没有说服力的。当人们根据数据发明某种理论时，如果这种理论没有合理的基础，没有得到新数据的检验，那么这种证据是无法令人信服的。神枪手应当仅仅画出一个目标，而且应当在开枪之前而不是之后画出这个目标。

输电线理论是否拥有任何依据？科学家对电磁场非常了解，但是没有任何合理的理论能够证明输电线电磁场会导致癌症。这种电磁场的能量远远低于月光的能量，其磁场能量比地磁场还要弱。

此外，费克廷和阿尔鲍姆在癌症得到诊断的那一年发现了癌症与儿童电磁场辐射量之间的相关性，但他们没有发现诊断一年前、五年前或者十年前癌症与儿童电磁场辐射量之间的相关性，尽管癌症症状的出现通常需要几年时间。最后，如果输电线电磁场与癌症之间存在联系，那么受到更高剂量辐射的人应当拥有更高的风险率，这与事实不符。他们自己的数据与他们的理论相矛盾。

这项瑞典研究得到新数据的证实了吗？没有。例如，覆盖英格兰、苏格兰和威尔士儿童的《英国儿童癌症研究》发现，接受电磁场辐射量较多的儿童患上白血病或其他癌症的可能性要小一些，尽管这种差异的观测值不具有统计说服力。此外，在这项瑞典研究公开后，人们对啮齿动物进行了一些实验研究，发现比输电线生成的电磁场大得多的电磁场对死亡率、癌症发病率、免疫系统、繁殖力和出生异常没有影响。

在权衡理论观点和经验性证据之后，美国国家科学院认为输电线不是一种公共健康威胁，没有必要资助进一步的研究，更不要说拆除输电线了。美国的一份顶级医学期刊同意这种观点，认为我们应当停止将研究资源浪费在这个问题上。就连瑞典那项研究的一位共同作者也做出了让步，认为没有必要进行进一步的研究，除非某种理论能够解释电磁场导致癌症的原因。

1999 年，《纽约客》发表了一篇题为《癌症聚集的迷思》的文章，含蓄地批判了保罗·布罗之前撰写的文章。

×如何轻松识破一本正经地胡说八道

数据聚集现象无处不在，甚至存在于随机数据之中。想要寻找某种解释的人一定会找到一种解释。不过，某种理论与数据聚集现象相符并不是一种具有说服力的证据。人们发现的解释需要言之有理，而且需要得到新数据的检验。

类似地，向足够多的目标发射足够多的子弹的人一定会击中某个目标。对数百种理论进行检验的人一定会发现至少支持一种理论的证据。这种证据不具有说服力，除非理论是合理的，而且能够得到新数据的证实。

如果你听到某些数据支持某种理论，那么在你相信之前，请回答两个问题。首先，这种理论是否合理？如果不合理，不要轻易相信胡言乱语是合乎情理的。其次，房子里是

否存在一位德克萨斯神枪手？宣传这种理论的人在提出理论之前是否查看了数据？他是否在选定他所宣传的理论之前对数百种理论进行了检验？如果你看到了冒烟的枪口，不要急于做出判断，应当等待这种理论接受其他数据的检验。

第 12 章
终极拖延

20 世纪 70 年代早期，美国统计协会赞助了一系列论文，以展示统计工具的广泛应用。加州大学圣迭哥分校社会学教授戴维·P.菲利普斯（David P. Phillips）撰写的一篇论文首先提出了下面这个具有挑逗性的问题：

> 在电影和某些类型的浪漫主义文学作品中，我们有时会看到临终时的场景，垂死之人努力不咽下最后一口气，只为看到某件特别的事情。例如，一位母亲可能会延缓死亡时间，直到分别已久的儿子从战场归来。这些小说中的壮举能否发生在现实生活中呢？

这显然是一个有趣的问题。我们许多人会把今天能做的事情拖到明天。不过，死亡也能像减肥和清理房间一样得到拖延吗？哪些证据可以用于判断死亡能否得到推迟呢？没有人收集孩子拜访临终父母的数据。即使有人收集，我们也无法知道即将到来的访问对父母的死亡时间是否具有任何影响。

菲利普斯是一个聪明人，以优异成绩获得了哈佛大学学士学位，并且获得了普林斯顿大学博士学位。他想到了一个巧妙的方法：调查

著名人物能否将死亡推迟到生日庆祝活动之后。对于老人来说，生日并不总是快乐的，因为它不断向人们提醒着时间的流逝。不过，菲利普斯认为，名人的生日可以得到大家的庆祝，因此是值得他们活着看上一眼的。关于名人死在生日或其他纪念日的逸事有很多。也许，他们为了纪念这些日子而推迟了死亡时间?

17 世纪的英国医生、科学家和作家托马斯·布朗爵士（Sir Thomas Browne）死于 10 月 19 日，那是他的生日。他似乎预言了这件事:“第一天应当成为最后一天，蛇尾应当刚好在这个时候回到它的嘴里，他们应当在出生日终结，这的确是一种惊人的巧合。”此外，布朗的朋友和仰慕者凯内尔姆·迪格比爵士（Sir Kenelm Digby）死于 6 月 11 日，那是他的生日，也是他在斯堪德隆战胜荷兰和威尼斯海军的日子。因此，理查德·法勒（Richard Farrar）写下了下面的墓志铭:

在 6 月 11 日，他出生和死去
并在斯堪德隆进行了勇敢的战斗;
生日、忌日和胜利日竟是同一天
真是一件少有的事情！

这两个死在出生日的人是个案吗？这种事情经常发生、比正常的巧合更加频繁吗？加州大学欧文分校教授吉尔伯特·盖斯（Gilbert Geis）从《美国名人录》中随机选择了 2250 个人，发现只有三个人死在生日那天，这比正常的随机情况还要少。盖斯挖苦地总结道:“也许，致力于科学的托马斯·布朗爵士希望提供一些数据，以支持他的理论。”

盖斯的数据没有证明人们能够把死亡推迟到生日之后。不过，一些名人死在了其他一些值得注意的日子上。美国第二任和第三任总统约翰·亚当斯（John Adams）和托马斯·杰斐逊（Thomas Jefferson）都是在 1826 年 7 月 4 日去世的，这一天是《独立宣言》签署 50 周年。5

年后，第五任总统詹姆斯·门罗（James Monroe）同样逝世于 7 月 4 日，他们是否将死亡时间推迟到了 7 月 4 日呢？这仅仅是巧合吗?

这很可能仅仅是一种巧合。美国第一任总统乔治·华盛顿（George Washington）死于 12 月 14 日。没有人记得这一天，因为它没有值得纪念的地方。它不是独立日、圣诞节、元旦、华盛顿的生日、他妻子的生日、他的结婚纪念日、任何战斗的周年纪念或者任何特别的日子。我们天生喜欢寻找模式，因此我们发现三位总统死在 7 月 4 日是一件值得注意的事情，但是我们忘记了没有死在 7 月 4 日的所有总统、将军以及其他美国领导人。死在 6 月 28 日的第四任总统詹姆斯·麦迪逊（James Madison）呢？如果其他三位总统能够死在独立日，为什么麦迪逊不能呢？他难道不能凭借强烈的爱国心将死亡推迟 6 天并死在 7 月 4 日吗?

美国有几千位名人，但一年只有 365 天——这意味着许多人将会死在特别的日子上，或者与其他人死在同一天。我们记住了“特别”的死亡日期，没有注意到其他死亡日期。某人恰好死在值得注意的日子上，这一事实是不值得我们注意的。它是无法避免的。

巧合的人物和日期只能证明我们花了许多时间寻找巧合，无法证明其他任何事情。要想进行有效的统计检验，我们需要事先表明，我们将查看一组明确的人物和一组明确的日期，看一看死在这些特别日期的人数是否多得出乎意料。要想检验菲利普斯的推迟理论，我们可以考察一群名人，看一看死在生日这天（或者生日过后不久）的人数是否多于死在临近生日时的人数。

死亡凹陷和尖峰

菲利普斯在某种程度上做到了这一点。他考察了《400 位美国名人》中具有已知死亡日期的逝者——这本书是《美国历史百科全书》

的一部分，后者是一本巨著，集结了美国最重要的史实。菲利普斯的结论是，著名人物常常能够推迟死亡时间，因为出生月之前几个月的死亡人数少于预期（“死亡凹陷”），而出生月和之后几个月的死亡人数则要多于预期（“死亡尖峰”）。这份关于名人生日和忌日的著名研究得到了许多统计学教材的引用，包括我自己的统计学教材。

在这项研究结束之后，菲利普斯和许多人合作发表了其他一些得到广泛宣传的研究报告，这些报告支持了他的理论：犹太人将死亡时间推迟到赎罪日之后；犹太人将死亡时间推迟到逾越节之后；中国老年妇女将死亡时间推迟到中秋节之后。也许，终极拖延的观点的确有些道理。

随后，人们产生了怀疑。在我的统计课程中，三名学生试图在学期论文中用不同数据集合复制菲利普斯的生日/忌日结论。没有人发现具有统计显著性的死亡凹陷和尖峰。图 12.1 显示了他们的联合数据。横轴上是死亡月减出生月的值，死在出生月之前的人是负值，死在出生月的人是零，死在出生月之后的人是正值。

死在出生月前一个月的人数稍稍小于预期，死在出生月之前第二个月和第三个月的人数则要远远多于预期，而且出生月和出生月之后并没有死亡尖峰。随机数据可以在大约一半的情形中表现出至少和图 12.1 中每个月实际死亡人数和预期人数之间差异一样大的差异。

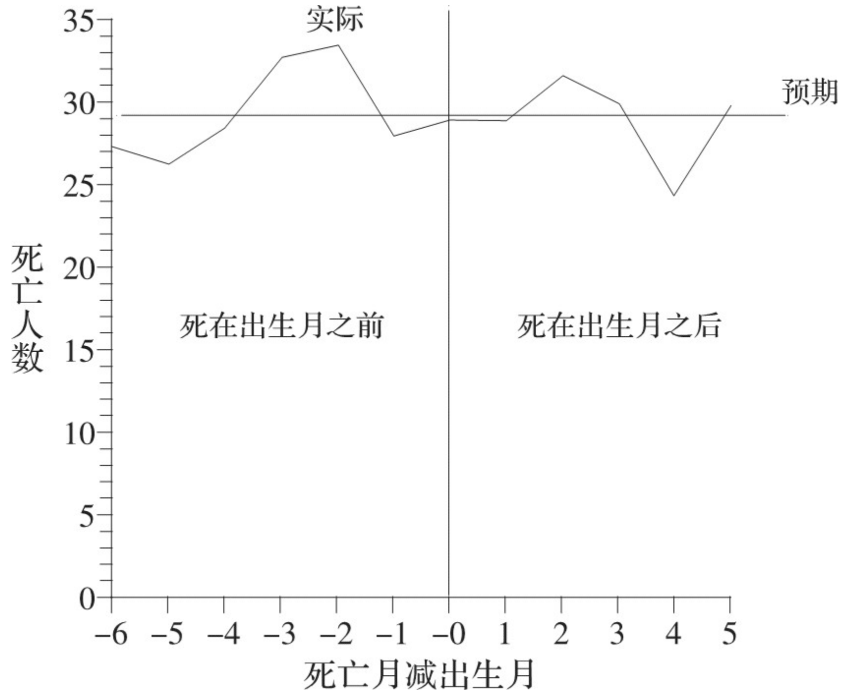


图 12.1　比较死亡月与出生月的三项研究

证实菲利普斯理论的三次独立尝试均告失败，这使人感到非常困惑，因此我和另一名学生希瑟·罗耶（目前已经当上了教授）重新检查了菲利普斯提出推迟理论所依据的生日/忌日数据。考虑到他对戏剧性临终场景的描述，他的分析中有一个令人不解的地方，那就是菲利普斯将出生月的所有死亡算在了一起，没有对生日之前和之后的死亡进行区分。而且，菲利普斯将出生月的所有死亡算作被推迟到生日庆祝之后的死亡。如果你生于 11 月 22 日，死于 11 月 11 日，他就会将你的情况算作成功的推迟。对于这种古怪的统计，唯一合理的解释是，这种做法可以使菲利普斯获得他想要的结果。

我和希瑟发现，在死于出生月的 26 位名人中，13 人死在生日之前，一个人死在生日当天，12 人死在生日之后！死亡日期与生日接近的 26 个人在死亡的推迟上完全没有取得成功。我们还将死亡时间分成了生日前 30 天，生日后 30 天……依此类推；我们并没有发现人们能够推迟死亡的证据。

经过进一步研究

这些发现使我对菲利普斯的其他研究产生了疑问。这些研究证明，人们能够将死亡推迟到赎罪日、逾越节和中秋节之后。这些研究也存在问题吗?

事实的确如此。我和另一名学生彼得·李（Peter Lee，现在是博士）考察了犹太人的数据。在赎罪日研究中，菲利普斯没有比较赎罪日之前和之后的死亡率。相反，他对赎罪日位于 9 月 28 日和 10 月 3 日之间的 10 年中（1922、1925、1930、1933、1941、1949、1952、1960、1963 和 1968）9 月死亡率与相邻两年的平均 9 月死亡率进行了比较。这是一种极为笨拙的考察赎罪日之前死亡凹陷的方法，而且完全忽略了节日之后的预期死亡尖峰。别忘了，这种理论认为，死亡推迟形成了之前的死亡凹陷和之后的死亡尖峰。要想支持这种理论，我们既要看到凹陷，也要看到尖峰。更加自然的做法是比较每年赎罪日之前和之后的死亡人数。不过，菲利普斯并没有提及这种明显的比较。更加奇怪的是，他的逾越节研究的确比较了逾越节之前一周的死亡人数和逾越之后一周的死亡人数。

生日、赎罪日和逾越节的研究使用了不同的方法：相邻月（并非以正式节日为中心），不同年份正式节日之前的一个月，以及相邻星期（以正式节日为中心）。这显然是一种可疑的做法。为什么一个人要在每项研究中使用不同的计算方法呢？答案似乎只有一个，那就是只有这样才能得到他所希望的结果。另一件可疑的事情是，赎罪日研究和逾越节研究使用了两组不同的数据。如果不是为了得到理想的结果，为什么不把这两组数据同时用在两个节日上呢?

问题还不止于此。令人震惊的是，研究中的死者不一定是犹太人！赎罪日研究中的数据包括 1921 ~ 1969 年纽约市的所有居民（这段时间犹太人口的比例为 28%）以及 1875 ~ 1915 年布达佩斯的所有居

民（这段时间犹太人口的比例为 22%）。逾越节研究中的数据包括通过姓氏判断“很可能是犹太人”的加利福尼亚死者。

根据生活在纽约或布达佩斯判断一个人的犹太人身份是一种极为勉强的做法。拥有看上去属于犹太人的名字不仅是一种勉强的判断标准，而且非常适合人为操纵。研究人员将名为阿舍和布罗迪的人包含进来，但却排除了名为阿什和巴赫的人。这是因为前者明显比后者更像犹太人，还是因为他们的死亡时间支持研究人员的理论？（突击测验：格林是犹太人还是非犹太人？）

彼得·李在一个犹太纪念教堂收集了超过 5000 名死者的数据，这些人不仅是百分之百的犹太人，而且具有强烈的信仰，因为他们的悼念仪式被安排在了犹太人停尸房里。我和彼得将他们的死亡日期与四个重要的犹太节日（光明节，逾越节，犹太新年，赎罪日）进行了比较。

总体而言，这些纪念日之前的死亡人数多于之后的死亡人数，尽管每个节日的结果都够不上统计显著性的阈值。对于每个节日来说，在随机情况下，观测到的差异出现的次数预计超过总次数的一半。

午夜的月饼

接着，我开始研究菲利普斯的中秋节数据。农历八月十五的中秋节是中国最重要的节日之一。这一天，一家人通常团聚在一起，庆祝节日。节日餐包括午夜时分在户外享用的传统月饼。

菲利普斯和另一位共同作者分析了加州死亡率，认为一些华裔美国老年妇女能够将死亡时间推迟到中秋节庆祝活动之后。最有力的证据是，他们观察到了 75 岁以上的中国妇女在中秋节之前一个星期的死亡凹陷以及之后一个星期的死亡增长。

这些数据的一个奇怪之处是，中秋节当天恰好拥有较多的死亡人数。这些死亡应当算作发生在节日之前还是之后呢？菲利普斯将它们算作节日之后；这样一来，它们就成了死亡推迟的证据。不过，主要的节日仪式发生在午夜。如果一个人能够将死亡推迟到节日之后，她难道不能将死亡推迟到主要的仪式活动之后吗？在逾越节前后的犹太人死亡研究中，菲利普斯将传统逾越节晚餐那天午夜之后的死亡算作成功的推迟。将发生在中秋节当天的死亡算作节日之后的唯一明显理由就是这样可以支持死亡推迟理论。

类似地，菲利普解释说，他之所以关注 75 岁的妇女，是因为老年妇女负责准备丰盛的晚餐。不过，在逾越节研究中，菲利普斯使用了 65 岁这一年龄作为分辨犹太老年男性的阈值。也许，中秋节研究使用 75 岁、逾越节研究使用 65 岁的原因是，只有这样做才能支持他的理论。

如果我们不是人为地将样本限制在“75 岁以上的妇女”这一范围内，不是莫名其妙地将节日当天的死亡视作午夜仪式之后发生的死亡，那么中秋节研究的结论可能不会具有统计显著性；这一结果完全不足为奇。

这些不合情理的现象要求我们使用没有经过必要篡改的新数据进行一项独立检验。我们用新的华裔美国人、韩裔美国人和越南裔美国人死亡数据对中秋节的死亡推迟进行了重新检验。不管如何记录节日当天的死亡，华裔老年妇女节日前一周的死亡人数都要多于节日后一周的死亡人数。越南裔老年妇女也是如此。韩裔老年妇女节日前的死亡人数要么多于节日之后，要么与节日之后持平。上述差异均不具有统计显著性。

这个例子真正的教训是什么？不是人们能够推迟死亡，而是如果你的目标是发表论文，而且你对数据的直接分析不起作用，那么你可以尝试不太直接的分析方法。做一个德克萨斯神枪手！迟早会有某种方法能够起到作用。进行大量检验，然后仅仅提及与你的理论相符的结果。

漫长的告别

在我进行这项“侦探”工作后不久，俄亥俄州立大学综合癌症中心统计学家和研究科学家多恩·扬（Donn Young）将他刚刚发表在《美国医学协会期刊》上的研究报告复印件寄给了我。他写道：

> 照顾临终癌症患者的医护工作者通常能够回忆起那些紧紧把握生命、排除万难活过重要节日或重要活动并在随后死去的人。通过“希望”自己活过某个节日，这些患者能够与家人共同经历这种重要活动，并且看上去为家人和朋友免除了将欢乐的节日和悲剧性事件联系在一起的不便。

为了弄清这些逸事是某种系统性现象的反映，还是对感人事件的选择性回忆，扬和一名同事考察了俄亥俄超过 30 万癌症患者的死亡时间。他们关注癌症患者的原因有三点。首先，同心脏病等疾病导致的突然死亡相比，死亡推迟与慢性病患者的关系似乎更加密切。其次，癌症患者很少与生命维持设备相连接，而连接这种设备的患者的死亡时间可以由家庭成员决定。第三，同其他许多疾病不同，癌症患者的死亡不具有季节模式，比如在冬季月份上升并在夏季下降。

他们比较了圣诞节、感恩节和患者生日之前以及之后几个星期的死亡人数。在所有三种仪式性活动中，活动之前的死亡人数稍稍多于活动之后的死亡人数，而且三种活动的差异均不具有统计显著性。他们的数据无法说明癌症患者能够将死亡推迟到重要活动之后。

多恩·扬将这些民间传说归结为美好的愿望和选择性记忆（我们记住了值得纪念的巧合，忘记了其他情况）。这种传说的一个不幸后果是，当病人死在节日之前时，家人可能会感到内疚或羞愧：约翰对我们爱的不够深，因为他不愿意和我们共同庆祝节日。扬给出了完全不

同的建议。如果你所爱的人得了绝症，而且某个重要活动即将到来，不要碰运气。现在就进行庆祝：

> 去年夏天，我的父亲在88岁生日的一个星期之后死于慢性肾衰竭，我们提前庆祝了他的生日，因为我们并不知道他能否坚持到这一天。幸运的是，他做到了，因此我们再次庆祝了他的生日。我觉得没有人会认为将这样的活动庆祝两次不是一个好主意。

×如何轻松识破一本正经地胡说八道

> 对于看上去天马行空的理论，应当抱有极为谨慎的怀疑态度。留意不自然的数据分组。留意研究人员似乎仅仅提到经过仔细选择的一部分统计检验的研究。

第 13 章
黑色星期一

1987 年 10 月 12 日的星期一是华尔街令人紧张的一个星期的开始，尽管这个星期最初看上去并不比平时更加紧张，而且没有什么使人感到紧张的特殊理由。接着，市场开始震荡。道指周三下跌 4%，周四下跌 1%，周五又下跌了 5%。纽约股票交易所（NYSE）创下了 3.39 亿股的交易纪录，输家与赢家的比例达到了 17:1。不过，人们很难弄清股价下跌的原因。

《华尔街日报》周一的“市场同步”专栏指出，周五的大跌可能是抛售的高潮，许多技术分析师相信，这是牛市开始的信号。一位分析师表示“抛售压力的峰值已经过去了”，另一位分析师表示“如果我们还没有看到底部，那么底部很可能近在眼前”。

如果股票市场的预测如此简单，那么这些专家就会通过股票交易发财致富，而不是依靠预测来谋生。这篇文章的刊登日期就是著名的黑色星期一，市场在这一天出现了可怕的动荡。当股市开盘时，由于沽盘订单数量众多，道指 30 只股票中的 8 只在纽约股票交易所的交易中断了一个小时。这一天的交易量达到了惊人的 6.04 亿股，道指下跌了 508 点（23%）。输家与赢家的比例达到了 40:1，所有股票的市值减少了大约 5000 亿美元。

第二天是“恐怖星期二”，股票市场接近全面崩溃。道指开盘时上涨了 200 点，但在中午之前回到了周一收盘点位以下。当天，许多著

名股票的交易出现了临时的中止，比如 IBM 中止了 2 个小时，默克中止了 4 个小时。一些大型金融机构据说即将破产，一些银行中止了对证券公司的借贷，证券公司要求之前借钱的交易人卖掉股票，偿还贷款。一些最大的证券公司敦促纽交所彻底关门，以阻止这种螺旋式下跌。

美联储当天宣布削减利率，承诺提供人们急需的资金，要求银行积极放贷；这些措施挽救了当天的股市。在投资银行的强烈鼓励下，一些大公司宣布了回购股票的计划，整个系统勉强脱离了解体的危险。一位市场参与者说："星期二是我们 50 年来经历过的最危险的一天。我想，再晚一个小时，股票市场就要崩溃了。"道指周二收盘时上涨了 6%，股票交易量为 6.08 亿，第二天又上涨了 10%——这宣告了短期危机的结束。

当少数数据与其他数据存在较大差异时，这些数据被称为"异常值"。例如，道琼斯工业股票价格平均指数每天的上下变动很少超过 4%。1987 年 10 月 19 日下跌的 23% 就是一个异常值。有时，异常值仅仅是笔误（比如弄错小数点的位置），而且可以得到校正。有时，异常值反映了特殊的情况，比如某一天发生自然灾害；如果我们不关心这种异常事件，我们可以将异常值丢弃。在其他情况下，异常值也许值得研究。我们可能非常希望看到股票市场对于自然灾害、总统遇刺或战争结束的反应。

1987 年 10 月 19 日黑色星期一的股市暴跌是一个值得研究的现象。其中的一个结论是，美联储愿意而且能够采取阻止股票市场暴跌所需要的任何行动。此外，总统特别工作组"布兰迪委员会"得到的结论是，所谓的"投资组合保险"（这个名称极具误导性，指的是在价格上升以后购买股票并在价格下降以后销售股票的投资策略）是导致 10 月 19 日暴跌和 10 月 20 日恐慌延续的元凶。布兰迪委员会还认为，股票、期权和期货市场在恐慌期间失去了联系，几乎进入了自由下落的状态。为避免此类事件再次发生，委员会建议由一个管理机构（比如美联储）对这些市场进行监管，为这些市场制定一致的借贷规则，并

在危机期间同时中止所有三个市场，以便为买家和卖家的匹配提供时间。这个异常值是一个痛苦而昂贵的教训——但它至少是一个教训。

在其他情形中，我们可以将异常值丢弃，因为它可能具有误导性。例如，一份国家级杂志曾经报道说，一群科罗拉多教师没能通过历史测试，其平均分只有 67 分。实际上，只有四位教师参加了这项测试，其中一个人得了 20 分。其他三个人的平均分是 83 分。一个非常低的分数将平均分拉低到了 67 分，误导了杂志，使其将平均分解释成了典型分数。在这里，异常值不具有代表性，而且会产生误导作用。

另一方面，丢弃异常值的做法有时会起到误导效果。一个戏剧性的例子发生在 20 世纪 80 年代，当时科学家发现，分析南极上空卫星臭氧读数的软件自动忽略了大量非常低的读数，因为同 20 世纪 70 年代得到的读数相比，这些数据属于异常值。计算机程序认为，这些与 70 年代正常值存在巨大差异的读数一定是错误的。

当科学家对这些数据（包括之前被忽略的异常值）进行重新分析时，他们发现，一个臭氧层空洞已经发展了几年，1985 年的南极臭氧比 1979 年减少了 40%。美国国家海洋和大气管理局高层大气实验室的苏珊·所罗门（Susan Solomon）说：“臭氧的这种变化程度绝对是前所未有的。我们从未见过南极目前正在经历的事情。”异常值有时是笔误、测量误差或意外，如果无法得到校正或忽略，它们会扭曲数据。在其他时候，异常值是最重要的观测值，比如臭氧读数。

排除异常值有时会产生误导作用，不排除异常数据则几乎一定会产生误导作用。

火箭科学

在“挑战者”号航天飞机 1986 年发射日期的前一天晚上，工程师们对推迟发射的问题进行了讨论。在将近十年时间里，工程师们一直在

担心橡胶密封圈的问题，这种密封圈用于接头密封，以容纳燃烧推进剂产生的高压气体。美国宇航局 1978 年的一份备忘录警告称，“为了防止热空气泄漏，导致灾难性失效，必须改变设计。”

一些密封圈曾在之前的飞行中失效，工程师担心这是因为密封圈在低温下会失去弹性。这一次，起飞时的温度预计将低于冰点，因此一些工程师反对发射。

图 13.1 显示了发生密封圈失效的 7 次飞行中的密封圈失效次数。他们并没有低于 53 华氏度[①]的数据。不过，在他们拥有的数据中，根据图 13.1，温度与密封圈失效次数之间没有明显的关系（除非我们能在理论上解释为何高温和低温会导致更多失效）。温度与密封圈性能之间缺乏明确关系的事实与人们不想打乱发射计划的强烈愿望压倒了一些工程师的担忧。

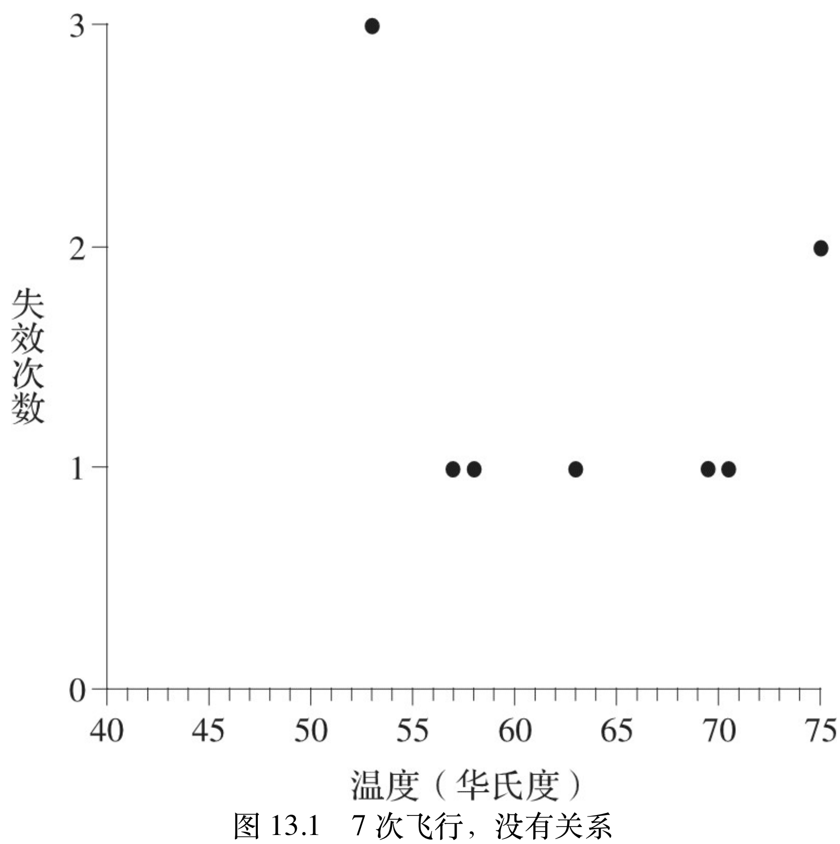


图 13.1　7 次飞行，没有关系

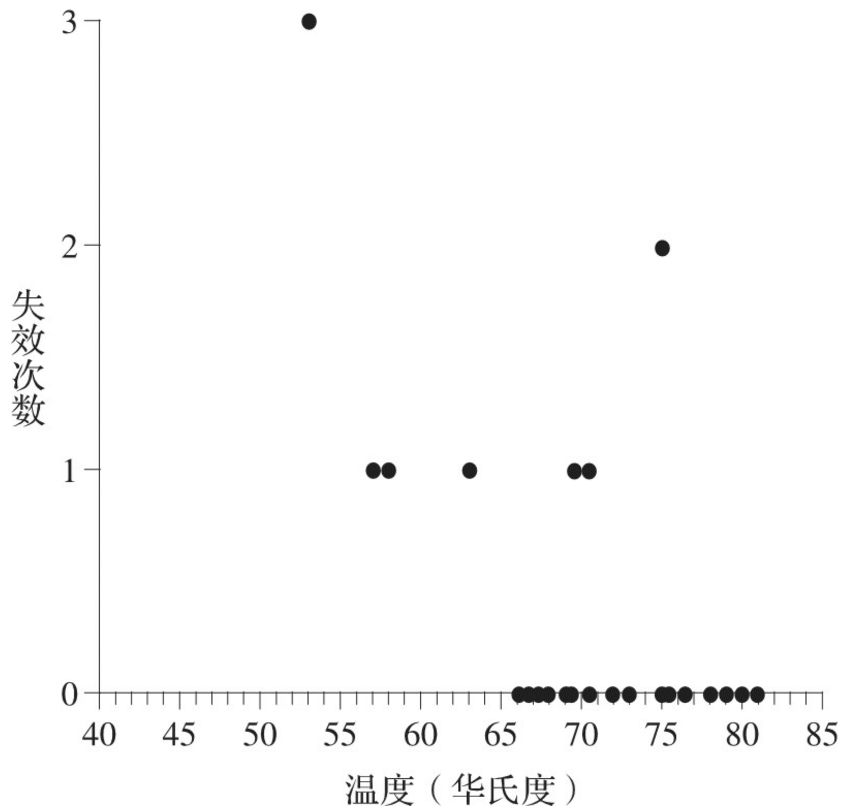


图 13.2　另一张图，包括所有 24 次飞行

图 13.1 的失败之处在于，它排除了没有发生密封圈失效的 17 次飞行。包含这些飞行的图 13.2 显示了完全不同的图景。在温度高于 65 华氏度的 20 次飞行中，只有 3 次飞行发生了密封圈失效。在温度低于 65 华氏度的 4 次飞行中，每次都发生了失效。更加不妙的是，之前失效次数最多的飞行具有最低的温度（53 华氏度），而此次飞行的温度预计将低于冰点。

不幸的是，工程师的担心是有道理的。发射当天，一个密封圈出了问题。73 秒钟后，“挑战者”号解体，7 名机组成员全部遇难。调查此次灾难的总统委员会中包括理查德·费曼，这位才华横溢的加州理工学院教授喜欢在露天酒吧里准备讲座。当他在洛斯阿拉莫斯的曼哈顿计划中开发原子弹时，他曾打开上锁的橱柜，并且通过篱笆上的洞偷偷溜出大院，以证明这个戒备森严的场所是不安全的。他的一本自传叫《别闹了，费曼先生》。

此时，费曼非常严肃。在全国电视转播的听证会上，这位曾经获得诺贝尔奖的理论物理学家将密封圈浸泡在一杯冰水中，从而简单地证明了密封圈在低温下会失去弹性。奇怪的是，火箭科学家们并没有想到这一点。这有点像弗朗西斯·培根（Francis Bacon）在 17 世纪讲过的一个寓言：

> 我主 1432 年，一群教友对于马嘴里的牙齿数量展开了激烈的辩论。愤怒的争吵整整持续了 13 天。人们翻出了所有古代书籍和编年史，并且表现出了这个地区之前从未见过的广博厚重的学识。到了第 14 天，一名举止优雅的年轻修士请他博学的前辈们允许他说上一句话。他早已对争论双方的智慧感到恼火和愤怒。随后，令人吃惊的是，他恳求他们以一种前所未闻的低俗方式取得和解：朝马儿张开的嘴里瞧上一眼，找到这个问题的答案。

在“挑战者”号的例子中，对于重要数据的忽略是一个无心而致命的错误。在其他情形中，人们故意忽略一些数据，因为这些数据不支持他们事先形成的观念。为了相信某件事情是正确的，他们丢弃了与这种信念相冲突的数据。

巴斯克维尔猎犬

在日语、普通话和广东话中，“四”和“死”的发音非常相似。许多日本人和中国人认为四是一个不吉利数字，这并不令人吃惊。真正令人吃惊的是，有人认为，由于这种强烈的厌恶，日裔美国人和华裔美国人很容易在每个月第四天遭到心脏病的袭击。这种说法显然很荒

谬。不过，一项得出这种愚蠢观点的研究报告发表在了一份世界顶级医学期刊上。

这份报告题为《巴斯克维尔猎犬效应》。在亚瑟·柯南·道尔爵士（Sir Arther Conan Doyle）的故事中，查尔斯·巴斯克维尔（Charles Baskerville）受到了一只恶犬的追逐，并且死于心脏病：

> 那只狗受到主人的激励，跃过边门，朝不幸的从男爵追去，从男爵则尖叫着沿着紫杉小径奔跑。在那条阴暗的通道上，那只牙齿和眼睛闪着火焰的巨大黑色生物跳跃着朝它的猎物追去，这幅景象一定非常可怕。由于心脏病和恐惧，从男爵在小路尽头倒了下来，没了气息。

我们每天都会看到“四”这个数字——在钟表、地址、电话号码、页码、价格、汽车里程表里。亚裔美国人真的如此迷信，如此惧怕每个月都会出现的第四天，感觉这一天就像在黑暗的小路上追赶自己的恶犬吗?

这项巴斯克维尔研究（它的缩写 BS[②]是不是很合适？）考察了死于冠心病的日裔和华裔美国人的数据。一种自然的检验方法是比较每个月第三天、第四天和第五天的冠心病死亡人数。在他们的数据中，在这三天的冠心病死亡病例中，33.9% 的死亡发生在每个月第四天，这与 33.3% 的期望并不存在巨大差异或统计上的差异。如果三号、四号和五号发生冠心病的可能性相等，那么这种程度的差异预计将在超过一半的情形中发生。

那么，巴斯克维尔研究是如何得到相反结论的呢？作者并没有提到 33.9% 这个数字。相反，他们提到了某些心脏病类型的死亡人数，同时并没有提到其他心脏病类型的死亡人数。在《国际疾病分类》中，冠心病被分成了许多类别。在一些类别中，四号的死亡人数超过三分之一。在其他类别中，四号的死亡人数不足三分之一。巴斯克维

尔研究仅仅提到了前一种情况的结果。他们丢弃了不支持这种理论的数据。

巴斯克维尔研究的首席作者与其他人共同进行了另外三项研究，其中两项研究使用了所有的心脏病类型，另一项研究使用了完全不同的心脏病类型。在不同研究中使用不同数据类型的唯一理由是，只有这样才能构造出支持结论的数据。

当我们怀疑某个研究人员在查看数据后进行了取舍时，我们可以尝试用新数据复制研究结果，以检验这种怀疑。巴斯克维尔研究使用了 1989 ~ 1998 年的数据。在使用 1969 ~ 1988 年以及 1999 ~ 2001 年的数据对巴斯克维尔研究中提到的心脏病类别进行重新检验时，得到的结果既不明显，也不具有统计显著性。在 1969 ~ 1988 年的数据中，五号的死亡人数多于四号；在 1999 ~ 2001 年的数据中，三号的死亡人数多于四号。另一个奇怪的现象是，作者可以使用 1969 ~ 1988 年的数据（他们在其他研究中使用了这些数据），但他们没有在巴斯克维尔研究中使用这些数据。我们可以猜到其中的原因。

支持这种贬低性理论的唯一证据是通过忽略不支持这种理论的疾病和年份得到的。如果没有这种有利的删减，就不会有“每月四日对亚裔美国人具有致命影响”的证据。这项研究真正证明的事情是，不管一种理论多么愚蠢，只要丢弃不支持这种愚蠢理论的数据，我们就可以找到支持它的数据。

五行噩运

中国农历的每一年与五行之一（金木水火土）相联系。传统中医认为，五行中的每一种元素与一种存储能量的脏器以及一种制造能量、移除废物的腑器有关（表 13.1）。

表 13.1　五行与人类器官

元素	脏器（阴）	腑器（阳）
火	心	小肠
土	脾	胃
金	肺	大肠
水	肾	膀胱
木	肝	胆囊

巴斯克维尔研究的首席作者与其他人共同进行了另一项具有争议性的研究。这一次，他们认为，中国人的出生年份可能会给他们带来“噩运”，即他们可能患上与这些年份相关的器官疾病。例如，1937 年是火年，火对应的脏器是心，因此 1937 年出生的中国人可能更容易死于心脏病。

也许，感觉自己终将死于心脏病的人不会努力预防这种疾病。或者，这些人的长期焦虑导致了他们所预想的结果。或者，二者之间也许没有任何关系。

作者报告说，华裔美国人的死亡数据支持他们的理论。不过，他们只提到了与表 13.1 中的 10 种器官相关的一小部分疾病。例如，他们提到了肾炎和另一种肾病的结果，但是没有提到其他肾病的结果；他们提到了支气管炎、肺气肿和哮喘的结果，但是没有提到其他呼吸系统疾病的结果。他们丢弃了不支持这种理论的数据。

作者将心脏病分成了不同寻常的子集，然后再次丢弃了碍事的数据。他们提出的理论是出生在火年的中国人更容易患上心脏病，不是出生在火年的中国人更容易患上某些类型的心脏病，不太容易患上其

他类型的心脏病。更加可疑的是，首席作者没有使用巴斯克维尔研究中的子集。他们显然对数据进行了考察，并在每项研究中选择了不同的子集，这种做法使统计检验变得毫无意义。

假设我宣布，我能在抛硬币时连续抛出正面，因为在我小时候，在其他孩子都在玩篮球和棒球的时候，我一直在练习抛硬币。（实际上，有一位离奇的魔术师兼数学家，名叫佩尔西·迪亚科尼斯，他能连续抛出正面。我没有这项技术。）不过，我可以制造证据证明我的抛硬币技能。假设我刚刚抛了 20 次硬币，得到了 9 次正面和 11 次背面。这没有任何奇特之处。不过，如果我能想出一种脆弱的借口，丢弃 10 个背面，那么，瞧，我在 10 次抛硬币实验中抛出了 9 个正面。这是令人信服的证据吗？当然不是。

当人们发表真正的实证研究（不是我的抛硬币实验）时，我们通常不知道他们是否丢弃了某些数据。一位研究人员可能会说，“我们使用了这些年份、这些人和这些疾病的数据”。看上去，这些类别像是提前选出来的。研究人员不会说，“在考察数据之后，我们丢弃了这些年份、这些人和这些疾病的数据，因为它们不支持我们的结论。”研究人员为什么要这样做呢？首先，他们希望得到能够发表的结果，而且觉得要想成功发表，他们需要得到统计显著性。其次，他们真诚地相信自己的理论是正确的，因此认为忽略相反的证据是没有关系的。

在噩运研究中，每一年对应于五行（金木水火土）之一。要想检验疾病与出生年份相关的理论，最自然的方法是考察死于某种疾病的中国人出生在噩运年份的比例是否超过 1/5。例如，既然火对应的脏器是心，那么死于心脏病的中国人出生在火年的比例是否超过 1/5？

既然作者没有提到这种检验，那就让我来介绍一下吧。在他们提到的 15 种疾病中，20.08% 的人出生在噩运年份。这既不是疾病与出生年份之间存在关联的明显证据，也不是具有统计显著性的证据。对于表 13.1 中更加广泛的疾病类别，19% 的人出生在噩运年份。这几乎具有了统计显著性，但方向是反的。

另一种合理的检验是考察在死于特定疾病的中国人中，出生在噩运年份的群体的寿命是否低于出生在其他年份的群体。在研究提到的15 种疾病中，只有两种疾病具有统计显著性。对于更加广泛的疾病类别，统计显著性消失了。例如，作者指出，出生在噩运年份并且死于支气管炎、肺气肿和哮喘的人寿命相对较短；不过，如果我们考察所有呼吸系统疾病，我们就会发现，出生在噩运年份的人拥有相对较长的寿命。

在没有被数据挖掘所污染的独立数据中，我们能够发现什么呢？1960～1968 年以及 1991～2002 年的数据无法复制作者对 1969 到 1990 年的数据进行搜刮以后得到的结果。对于他们声称受到出生年份影响的疾病，几乎刚好有 20% 的死者出生在噩运年份，而且出生在噩运年份和其他年份的亚裔美国人不具有明显或具有统计显著性的平均死亡年龄差异。

和巴斯克维尔研究一样，出生年份噩运研究真正能够证明的结论是，对于任何理论，只要考察大量数据，并且丢弃不支持这种理论的数据，那么你一定能够收集到支持这种理论的证据。

花押决定论

“我们的喜悦之杯正在满溢！我们有了一个女儿，她有着仁慈的大自然所能给予人类的最为匀称的比例和最为可爱的仪表，她的名字叫做伊玛（Ima）！”人们之所以为这个天使般的婴儿选择这个名字，是因为她的叔叔写了一首宏大的南北战争叙事诗，诗里的女主角叫做伊玛。这份出生声明是由伊玛热情的父亲詹姆斯·史蒂芬“大个子吉姆”·霍格（James Stephen “Big Jim” Hogg）起草的。我们并不知道这位父亲是否意识到他的女儿获得了一个不幸的名字：伊玛·霍格[③]。

伊玛后来说，“我的祖父斯廷森住在距离米尼奥拉 24 公里的地方，当时消息传播得很慢。当他听说孙女的名字时，他用最快的速度跑到镇上进行抗议，但是已经晚了。洗礼已经结束了，我只能使用伊玛这个名字。”

大个子吉姆后来意识到了女儿名字的幽默之处。德克萨斯长期流传着这样一个故事：大个子吉姆将这种幽默用在了政治上，他在重新竞选州长的演讲中对人群说，他有两个女儿，分别是伊玛和乌拉。《堪萨斯城星报》为这个故事添加了更多细节，说他还有第三个女儿，名叫胡莎。实际上，伊玛没有姐妹，只有三个名字普通的兄弟：迈克尔、托马斯和威廉。

虽然拥有滑稽的名字，但伊玛·霍格活得很好。大个子吉姆买下了德克萨斯州西哥伦比亚附近的土地，他在遗嘱中说，这片土地在他去世 15 年以后才能卖给别人。在吉姆去世 12 年以后，这片土地发现了石油，伊玛和她的兄弟从一般富裕变成了极度富裕。伊玛热爱德克萨斯，并且开展了广泛的慈善活动，因此获得了“德克萨斯第一小姐”的称号。

除了婴儿的名字，家长是否还应该考虑他们的首字母缩写呢？我的首字母缩写是 GNS（加里·南斯·史密斯）；我弟弟的首字母缩写是 RAS（罗伯特·艾伦·史密斯）。如果我的父母将我命名为加里·艾伦·史密斯呢？GAS 这个首字母缩写带来的羞愧感会毁掉我的人生吗？我会通过某种方式战胜这种耻辱吗?

加州大学圣迭戈分校（UCSD）心理学和社会学系的一群研究人员认为三个字母的姓名缩写是非常重要的，因为同拥有消极姓名缩写（比如 PIG 或 DIE）的人相比，拥有积极姓名缩写（比如 ACE 或 VIP）的人寿命比较长。这种寿命受首字母缩写影响的观点被称为“花押决定论”。奇怪的是，加州大学圣迭戈分校的研究人员还发现，消极的姓名缩写比积极的姓名缩写更加常见，而且比过去更加普遍。难道家长具有虐待倾向，而且这种倾向正在变得越来越严重吗?

如果消极的首字母缩写对健康具有如此巨大的危害，那么它似乎不太可能变得如此流行。慈祥的父母当然不会选择那些能够缩短孩子预期寿命的名字。这项研究一定存在某些问题。

加州大学圣迭戈分校的研究人员使用了 12 种积极的首字母缩写（ACE，GOD，HUG，JOY，LIF，LIV，LOV，LUV，VIP，WEL，WIN 和 WOW）以及 19 种消极的首字母缩写（APE，ASS，BAD，BUG，BUM，DED，DID，DTH，DUD，HOG，ILL，MAD，PIG，RAT，ROT，SAD，SIC，SIK 和 UGH）。他们请大学生将这些首字母缩写以及 9 种中性缩写标记为积极、消极和中性缩写，以证实这种分类方法。

我们不知道研究人员看到数据时是否已经选择了这些首字母缩写。虽然学生能够分辨缩写的好坏，但这并不能告诉我们这份清单是如何制定的。要想复制这项研究，我们可以从包含 100 个积极或消极首字母缩写的独立清单入手，其中包括圣迭戈分校研究人员使用的 31 个缩写。然后，我们可以将这份清单提供给一组学生和教员，请他们选出最好和最坏的缩写。

表 13.2　排在最前面的积极和消极缩写，按顺序排列

积极	消极	
ACE	ASS	GAS
ICE	KKK	FAT
JOY	FAG	BAD
VIP	DIE	POX
CEO	GAY	HOR

WOW	ZIT	BUM
GEM	FUK	SIN
FLY	PIG	
FOX	DUM	
HIP	RAT	
WIT	SOB	
WIN	TIT	

表 13.2 显示了最好的 12 个积极缩写和最不好的 19 个消极缩写。在圣迭戈分校研究人员的 12 个积极缩写中，只有 5 个出现在了前 12 个最好的缩写中（用粗体显示）；在研究人员的 19 个消极缩写中，只有 6 个出现在了前 19 个最不好的缩写中。我们很难解释其中的一些遗漏，尤其是消极的缩写。也许，这些缩写是在他们查看数据之后选择的?

为什么他们不为积极和消极缩写选择相同的数量，而是选择了 12 个积极缩写和 19 个消极缩写？为什么他们使用了 BUG 而不是 FAG、DUD 而不是 DUM、HOG 而不是 FAT？他们显然丢弃了不支持这种理论的首字母缩写。

圣迭戈分校研究人员使用了 1969 ~ 1995 年的死亡数据，根据死亡时间将死者分成了不同小组，并将拥有积极或消极首字母缩写的群体的平均死亡年龄与拥有中性首字母缩写（这些缩写可能也是在查看数据之后选择的）并且死在同一年的群体进行了比较。他们报告说，平均而言，拥有积极首字母缩写的男性多活了 4.5 年，拥有消极首字母缩写的男性少活了 2.8 年，拥有积极首字母缩写的女性多活了 3.4 年，拥有消极首字母缩写的女性没有区别。4.5 年的预期寿命差异超过了美

国和委内瑞拉之间的差异，几乎相当于美国和阿尔及利亚之间的差异。美国同委内瑞拉和阿尔及利亚之间的差异存在合理的解释。首字母缩写对预期寿命可能具有的如此巨大的影响则没有同样科学的解释。

根据死亡年份对人们进行分组的做法存在一个问题，那就是不同时期各种首字母缩写的普遍程度是不同的。例如，假设拥有积极和消极首字母缩写的人具有相同的死亡率，但是现在的消极首字母缩写比过去更加普遍（这是事实）。在这种情况下，最近具有消极首字母缩写的死者就会比具有其他首字母缩写的死者更加年轻。

举一个极端的例子。假设首字母缩写 ACE 在 50 年前不再被人使用，首字母缩写 GAS 则在同一时间开始被人使用。这显然是不现实的，但它是对“消极首字母缩写正在变得更加普遍”这一现象做出的便于理解的夸张。在我们这个模式化例子中，所有首字母缩写为 ACE 并且死于去年的人在去世时至少有 50 岁，所有首字母缩写为 GAS 并且死于去年的人在去世时不到 50 岁。这种差异与首字母缩写对预期寿命的影响没有任何关系。它仅仅与不同首字母缩写普遍程度的变化有关。

如何规避这个问题呢？我们可以按照人们的出生年份对他们进行分组。60 年前出生的、首字母缩写为 ACE 的人应当与 60 年前出生的、首字母缩写为 GAS 的人进行比较。40 年前出生的、首字母缩写为 GAS 的人应当与 40 年前出生的、具有其他首字母缩写的人进行比较。这样一来，我们所观测到的任何死亡年龄差异将来自预期寿命差异，而不是不同首字母缩写流行程度的变化。

实际上，当我们对同一年出生的人进行比较时，首字母缩写与死亡率之间没有任何关系。平均而言，同具有积极首字母缩写的男性和女性相比，具有消极首字母缩写的群体的寿命要稍微长一些（这与圣迭戈分校研究人员得到的结果相反），但是这些差异很小，不具有统计显著性。

我还试图用时间范围更长的数据（1905～2003 年）对圣迭戈分校的研究进行复制。这些数据同样无法支持他们的结果。不管使用这些研究人员的首字母缩写清单还是通过调查得到的缩写清单，平均而言，具有消极首字母缩写的男性寿命要稍微长一些，女性的结果则恰好相反，但是这些差异的观测值既不明显，也不具有统计显著性。

×如何轻松识破一本正经地胡说八道

在研究中遗漏数据的做法是一个巨大的危险信号。包含或者排除数据的决定有时具有天壤之别。这种决定应当基于数据的相关性和质量，而不是数据是否支持或影响研究人员期望或理想的结论。

对于得到错误记录的数据进行校正是可以的。有时，忽略异常值也是可以的。不过，对于研究人员来说，最好的规则是，当存在疑问时，不要排除数据。对于读者来说，最好的规则是对丢弃数据的研究保持警惕。问问你自己，被忽略的数据是否存在非常明显的错误。如果不存在，说明研究可能有问题。一些数据之所以被丢弃，可能仅仅是因为它们与理想的结果相矛盾。

① 摄氏度(°C)=［华氏度(°F)-32］÷1.8——译者注

② BS 意为胡说。——译者注

③ 与“我是猪”谐音。——译者注

第 14 章
点球成金

对统计学家来说，美国职业棒球大联盟（MLB）是一个极具吸引力的事物，因为它拥有超过 100 年的详细资料。更加美妙的是，一个名叫肖恩·拉赫曼（Sean Lahman）的棒球迷收集了美职棒历史上所有选手和队伍的大量统计数据，并且允许所有人下载这些数据。让我们开始搜刮数据吧！

许多统计学家（包括具有传奇色彩的比尔·詹姆斯）提出了赛伯统计量（也就是基于数据的、对选手和球队表现的客观测量），彻底改变了我们对棒球的理解。赛伯统计学家提出和支持的许多统计量现在已经成了人们司空见惯的事物。

例如，一名选手的击球技能传统上是用安打率衡量的，即安打数量除以击球次数，不包括保送、触身球以及牺牲长打。赛伯统计学家认为安打率无法很好地衡量击球手对球队得分的帮助作用，后者才是真正重要的事情。安打率没有考虑自由上垒和触身球的情况，而且没有对一垒、二垒、三垒和全垒打进行区分。

赛伯统计学家提出的另一个指标是上垒加长打率（OPS），其中上垒平均分包括安打、保送和触身球，长打平均分则是对垒打数的统计（一垒为 1 分，二垒为 2 分，三垒为 3 分，全垒打为 4 分）。上垒加长打率不是一种完美的指标，但它很容易被人理解，而且相对于传

统安打率是一种巨大的进步。目前，这种指标经常出现在新闻报道和棒球卡片中，甚至被用在少年棒球联合会的选手身上。

迈克尔·刘易斯（Michael Lewis）的畅销书《点球成金》宣传和肯定了赛伯统计量，这本书描述了奥克兰运动家队用赛伯统计量发现被低估的（即便宜的）选手，以便与薪金总额高得多的球队竞争的故事。如果另一支球队认为安打率远高于胡安的鲍勃更加优秀，而奥克兰队认为上垒加长打率更高的胡安比鲍勃更加优秀，奥克兰队就会尝试用鲍勃交换胡安。具有讽刺意义的是，目前，许多美职棒球队都在使用赛伯统计量，因此奥克兰队的优势受到了削弱甚至逆转。

对学术界人士来说，棒球统计数据的宝库为他们提供了一个完全不同的机遇——他们有机会在数据中发现具有统计显著性、因而可以写进论文之中的模式，尤其是当他们对数据进行微调时。在这些被人发现的模式中，许多模式几乎没有任何意义，但它们还是得到了发表，而这仅仅是因为它们具有统计显著性。

棒球迷信

用一根圆圆的木棒击打一个以每小时 144 公里的速度飞行、可能具有左旋、右旋、上旋或下旋的圆形棒球也许是一切运动项目中最为困难的挑战。如果球被击中，它可能直接飞向外野手，造成出局，或者安全落地，形成安打。平均而言，棒球选手每四次击球只能击出一次安打。如果你能将这个比例提高到 1/3，你就能进入名人堂。从投手的角度看，微小的差异足以区分所有明星投手、熟练投手和失败者。

也许，这就是棒球选手的迷信如此出名的原因，他们总是在寻找有可能使机会的天平倒向自己一边的某种事物——不管这种事物多么可笑。奇怪的是，研究人员常常会做同样的事情，寻找看上去具有某种模式的事物——不管这种事物多么可笑。

三垒手韦德·博格斯（Wade Boggs）在职业生涯中的表现非常出色，他主要在波士顿红袜队打球。他的职业生涯安打率为0.328（1/3）。他曾经连续12次参加全明星赛，并在2005年入选棒球名人堂。如果这个世界上有一个迷信名人堂，那么他一定也会入选。博格斯每天在完全相同的时间起床，并在下午2:00吃鸡肉。他以14天为一个周期，轮换使用13种食谱（包括两次柠檬鸡）。当他晚上需要到芬威公园参加比赛时，这位鸡肉男会在3:30准时进入更衣室，穿上队服，并在4:00去球员席热身。他会进行精确的热身程序，包括接150个地滚球。在防守热身结束时，他会站上三垒、二垒、一垒以及垒线（当他入场比赛时，他会跳过垒线），用两步走到教练席，用四步走到球员席。赛季结束时，博格斯的步伐在草地上留下了永久性的脚印。博格斯总是在5:17进行击球练习，并在7:17进行冲刺练习（一位反对他的经理曾经让体育场的时钟从7:16直接跳到7:18，以便矫正他的行为）。

比赛期间，当博格斯走上三垒时，他会用左脚清理面前的泥土，拍三下手套，正一正帽子。每当他击球时，他都会在击球区画上意为“生命”的希伯来语单词。（他不是犹太人。）

虽然博格斯非常迷信，但同特克·温德尔（Turk Wendell）相比，他几乎是正常的。从1993年到2004年，温德尔在美职棒做了11年投手。他身穿99号队服，因为这是电影《大联盟》中查理·辛（Charlie Sheen）扮演投手里基·沃恩（“野兽”）时使用的球衣号码。当温德尔2000年与纽约大都会队签约时，他所要求的合同金额为9999999.99美元。他戴着一条项链，那是用他所猎杀的野生动物的牙齿制作的。在走上投手丘之前，温德尔会把四颗黑色甘草糖放进嘴里，以便在投球时咀嚼。他不是踩过垒线（一些选手会这样做）或者跨过垒线（另一些选手会这样做），而是跳过垒线。当他站上投手丘时，他会朝中外野手挥手并等待对方向他挥手。当他需要新的棒球时，他会要求裁判员把球滚过来，因为他不想捕球；如果裁判员把球扔给他，温德尔要么让球打在他的胸前，要么将球放走。每局比赛结束后，他会在球员

席刷牙，这种不同寻常的举动甚至印在了厄珀代克公司的棒球新人卡上。

棒球迷信也适用于球队，其中最著名的是“班比诺诅咒”。波士顿红袜队 1918 年赢得了第五次世界职业棒球大赛冠军，成为了当时最成功的棒球队。在 1920 年赛季开始之前，红袜队老板将球队最佳球员贝比·鲁斯（“班比诺”）卖给了当时还没有获得过世界职业棒球大赛冠军的纽约扬基队，得到了 12.5 万美元现金和 30 万美元贷款，用于资助由老板女友担任主角的百老汇音乐剧。

在扬基队的第一个赛季，鲁斯击出了 59 个全垒打，超过了其他任何球队的全队全垒打数量。此后，扬基队获得的世界职业棒球大赛冠军数量超过了其他任何球队，红袜队则一年比一年令人失望，迟迟无法获得冠军。他们四次进入世界职业棒球大赛，均铩羽而归——而且总是以戏剧性的方式在决定胜利的第七场比赛中失利。最后，在 2004 年，经过 86 年的失望，红袜队终于赢得了世界职业棒球大赛冠军。

芝加哥小熊队还没有打破他们的诅咒。芝加哥小熊队在 1907 年和 1908 年赢得了世界职业棒球大赛冠军，此后再也没有获得这项荣誉，形成了长达 100 多年的冠军荒。据传说，这个诅咒与一只山羊有关。这只山羊从一辆卡车上掉下来，走进了芝加哥一个酒吧。为了吸引顾客，酒吧主人比利·塞亚尼斯收留了这只山羊，留了一副山羊胡子，开始自称比利·山羊，并将酒吧的名字改成了比利山羊酒吧。

在 1945 年世界职业棒球大赛小熊队对底特律老虎队的第四场比赛中，比利买了两张包厢票。当时，小熊队以总比分 2:1 领先。比利带着山羊来到了球场，他披着一条毯子，上面写着“我们得到了底特律的山羊”。

小熊队老板菲尔·里格利（Phil Wrigley）对员工说：“那只山羊很臭。”于是，塞亚尼斯被要求带着山羊离开球场。塞亚尼斯在走的时候喊道：“小熊队再也赢不了了！”小熊队输掉了第四场比赛以及接下来的两场比赛，从而输掉了世界职业大赛。塞亚尼斯给里格利发了一封简短的电报：“现在谁臭？”

为了扭转这个诅咒，小熊队球迷将活山羊和死山羊带进里格利球场，并将山羊捐献给发展中国家的人民。不过，这些方法并没有产生任何效果。[①]

棒球迷信通常具有娱乐性质，不会带来任何危害。不过，韦恩州立大学两位多产的研究人员欧内斯特·艾贝尔（Ernest Abel）和迈克尔·克鲁格（Michael Kruger）指出，一些迷信可能是致命的。

如果你给我一个D，我就会死（die）

一些研究发现，人们（包括当事人在内）认为教名不受欢迎的人比不上教名受欢迎的人。如果健康与自尊有关，那么教名不受欢迎的人可能具有更低的预期寿命。不过，对于这个问题的一项详细研究发现，名字的受欢迎程度与预期寿命之间没有关系。

认为预期寿命可能受名字第一个字母影响的想法更加没有道理。一项对于工商管理硕士学生的小型研究发现，教名或姓氏以C或D开头的学生倾向于拥有较低的学分积。不过，在满分为4分的评分体系中，以C或D开头的学生的平均分只比其他学生低0.02。我们并不知道这个结果背后是否隐藏着大量没有被提及的检验。

不过，在这项脆弱结论的启发下，艾贝尔和克鲁格发现，在职业棒球大联盟中，同教名首字母为E到Z的选手相比，教名始于字母D的选手平均少活1.7年。原因何在？“众所周知，职业运动员非常迷信。”这的确是事实，但是这种模式化印象指的是选手的迷信行为，比如每场比赛穿同一双（没有洗过的）袜子，或者在入场时触摸一垒。穿着没有洗过的袜子和对名字的第一个字母感到不安是不一样的。真的有人相信一个人会因为名叫唐纳德（Donald）而不是罗纳德而早死吗？

关注字母 D 的做法并没有令人信服的理由。艾贝尔和克鲁格认为“D 不属于 ABC，而且几乎等同于失败。”几乎失败至少不像真正的失败（failure）那样糟糕，而后者通常是用 F 表示的。而且，字母表中有 23 个字母不属于 ABC，其中可能有许多字母看上去不太理想。

百事可乐做过一项单盲品尝试验，他们在标有 M 的杯子里装入百事可乐，在标有 Q 的杯子里装入可口可乐，然后让人们品尝。超过一半的人更喜欢百事可乐。随后，可口可乐也进行了一项试验，在标有 M 和标有 Q 的杯子里装入可口可乐，然后让人们品尝。他们发现，大多数人更喜欢标有 M 的杯子里的可口可乐。因此，他们的广告标题是：“可口可乐战胜可口可乐的日子”。显然，Q 是一个缺乏吸引力的字母。

在 26 个字母中，我们总能找到一些统计模式，尤其是当我们对数据进行编排并分析其中的一些子集时。随后（在发现事实以后），我们总能想到这些巧合关系的某种解释。如果 F 的结果不好，这是因为 F 代表失败。如果 X 的结果不好，这是因为 X 是一个不经常出现的字母。如果 Z 的结果不好，这是因为 Z 是字母表中的最后一个字母。要想进行有效的统计检验，应当在检验之前而不是之后确定这些“坏”字母。

学分积研究使用了 A、B、C、D 和 E～Z 五个类别，并且考虑了每个人的教名和姓氏首字母，排除了首字母存在冲突的人。戴维·琼斯（David Jones）和琼斯·戴维被分到 D 组，艾伦·戴维（Allen David）和戴维·艾伦则被排除在了分析之外。只有教名和姓氏的首字母都是 E～Z 的人才会被分到 E～Z 组，比如伊桑·弗莱明（Ethan Fleming）。

艾贝尔和克鲁格认为工商管理硕士的研究是导致他们开展这项研究的契机，但他们在自己的研究中只考虑了教名的结果。艾伦·戴维被分到 A 组，戴维·艾伦被分到 D 组，詹姆斯·戴维（James David）被分到 E～Z 组。而且，艾贝尔和克鲁格的结果只涉及 1875 年到 1930 年出生的人。他们没有提供将分析限定在字母 D、教名以及这种时间段上的原因，除了……这种限定可以得出他们想要的结果。

这些假设并不是随意选择的。艾贝尔和克鲁格得出的“名字始于字母 D 的棒球选手寿命较短”的结论取决于下列人为限制：使用教名而不是教名和姓氏，只考虑 1875 年到 1930 年出生的选手。否则，名字始于字母 D 的美职棒选手的平均死亡年龄差异将不具有统计显著性。

糟糕的出生月份

由迪士尼 1958 年拍摄的一部曾经赢得奥斯卡奖的纪录片展示了数千只旅鼠成群结队来到悬崖边跳海自尽的场景。解说员庄重地评论道：“每只小小的啮齿动物被一种迫切的力量攫住，被一种缺乏理性的歇斯底里驱使，步调一致地向前进发，走向一种奇怪的宿命。”这种戏剧性的场景使“旅鼠式的行为”成了美国语言的一部分，用于描述相互模仿、不关心后果的人。

旅鼠是一种啮齿动物。对于迪士尼这个将老鼠作为核心元素的公司来说，这部纪录片竟然是完全荒谬的，这是一种巨大的讽刺。旅鼠喜欢独处，不喜欢与水接触，而且可以在迫不得已的时候游泳。那么，迪士尼是如何拍摄旅鼠自杀之旅的呢？他们将几十只旅鼠放在被雪覆盖的转盘上，从不同角度拍摄它们在转盘上奔跑并被抛向空中的场面。经过编辑，影片形成了数千只失去理性的旅鼠争相赴死的效果。这种人为编排的场景竟然被称为纪录片！虚拟世界比现实更加奇怪。

旅鼠不会自杀，更不会大规模自杀。所有生物都有生存的本能。不过，人类的确会自杀。怎样的绝望环境会战胜他们生存下去的自然力量呢？

自杀的人显然认为他们的生命极其压抑而且毫无希望，因此做出了停止生存这个不可逆转的决定。自杀往往在春天和初夏更加常见——这也许是因为当天气转暖、百花盛开、爱情绽放之时，一些人被

自己的抑郁和其他人的幸福之间的反差所击溃。这是合乎情理的。不过，艾贝尔和克鲁格（认为名字以 D 开头的人会早死的那两个研究人员）提出了一个荒谬的观点：自杀倾向取决于一个人的出生月份——而不是他们选择死亡的月份。如果这是真的，它将是一个不同寻常的发现；可惜，它是假的。这项研究的数据和统计检验都是错误的。

表 14.1　棒球选手各个出生月份的自杀人数

	观测到的数字	**调整后的数字**	**期望值**
1 月	6	9	10.5
2 月	7	13	10.5
3 月	5	8	10.5
4 月	5	10	10.5
5 月	5	9	10.5
6 月	6	11	10.5
7 月	2	3	10.5
8 月	19	29	10.5
9 月	5	8	10.5
10 月	7	11	10.5
11 月	3	5	10.5
12 月	6	10	10.5

总计	76	126	126

艾贝尔和克鲁格考察了职业棒球大联盟自杀球员的出生月份。表14.1中的“观测到的数字”一列显示了他们的数据。7月出现了明显的下探，8月出现了明显的尖峰，随后的9月出现了下降。因此，艾贝尔和克鲁格认为，出生在8月的人拥有较强的自杀倾向。

一些出生月份将不可避免地比其他月份拥有更多的自杀人数。在谴责8月份之前，我们需要考虑研究人员观测到的这种模式是否合理，是否具有统计说服力。对于第一个问题，没有可靠的理由能够解释为什么出生在8月的人比出生在7月或9月的人拥有更高的自杀倾向。

要想进行统计分析，我们可以根据每月的出生人数对每月观测到的自杀人数与期望值进行比较。例如，如果所有选手中的10%的人出生在一月，那么我们预计10%的自杀者出生在1月。

艾贝尔和克鲁格采取了完全不同并且完全错误的做法。他们用每个出生月观测到的自杀人数除以出生在这个月的选手总人数，然后乘以1000，并且四舍五入到最近的整数，得到所谓的“调整后的数字”。这种令人困惑的计算结果（表14.1“调整后的数字”一列）将自杀总人数提高了66%，从76人提高到了126人。接着，他们用126除以12，得到了每个月的期望值10.5。完成这些操作以后，他们指出，出生月份与自杀之间存在具有统计显著性的关系。

且慢！我们不能编造出66%的自杀人数。要想得到更多数据，唯一有效的方法就是收集更多数据，比如将其他职业运动员包含进来。另一个问题是，他们的数据是错误的！在他们撰写这篇论文时，拥有已知出生月份的自杀选手为86人——不是他们所说的76人。而且，有5个自杀者出生在7月（比艾贝尔和克鲁格的数据多3个），有16个自杀者出生在8月（比他们的数据少3个）。他们将3个出生在7月的人统计成了出生在8月的人。

在校正这些错误以后，每月自杀人数差异的观测值不具有统计显著性。事实上，根据预期，我们可以在超过一半的情形中得到像观测值那么大的出生月份差异。

和以前一样，最后的结论是，在自杀出生月中，7 月下探、8 月激增、9 月下跌的说法没有任何理论基础；数据得到了错误记录；统计程序也是错误的。当正确数据得到正确分析时，出生月份与自杀之间没有任何关系。

名人堂的死亡之吻

在另一篇论文中，艾贝尔和克鲁格认为，“棒球名誉可能会带来沉重的代价”，因为入选名人堂的职业棒球大联盟球员的平均死亡年龄比其他球员低 5 岁。这种令人吃惊的论断与“积极的自尊可以增进身体健康”这一普遍观念相矛盾。实际上，艾贝尔和克鲁格曾经提出，教名首字母为 D 的棒球选手由于缺乏自尊而具有较短的寿命。如果像首字母为 D 这样微不足道的事情都能导致死亡时间提前，那么像成为行业佼佼者这样实实在在的事情当然应该延长预期寿命。

他们的研究一定有问题。事实的确如此。

艾贝尔和克鲁格的数据来自拉赫曼的数据库。对于入选名人堂时仍然健在的每个人，艾贝尔和克鲁格找出了在入选当日健在并且与入选名人堂选手出生在同一年的球员。他们记录了每个选手的死亡年份。如果数据中没有列出死亡年份，他们认为选手目前仍然健在。接着，对于样本中仍然健在的一些成员，他们使用了某种专门设计的统计程序。到目前为止，情况还算不错。

问题是，对许多人来说，拉赫曼的数据之所以没有列出死亡年份，不是因为选手仍然健在，而是因为死亡年份是未知的。每个入选名人堂的选手要么健在，要么拥有已知的死亡日期。相比之下，许多

不出名的选手并没有明确的死亡日期，尤其是棒球运动早期的选手。艾贝尔和克鲁格将死亡日期未知的过世选手视作仍然健在的处理方法，人为提高了他们所得到的没有入选名人堂的选手寿命。

例如，1937 年入选名人堂的赛·扬（Cy Young）生于 1867 年，死于 1955 年，终年 88 岁。为了进行对比，艾贝尔和克鲁格使用了出生于 1867 年并在 1937 年赛·扬入选名人堂时明显健在的 22 名选手。在 22 名选手中，19 人拥有已知的死亡日期，他们的平均死亡年龄是 83 岁，比赛·扬小 5 岁。3 名选手没有死亡日期，因此艾贝尔和克鲁格认为这 3 名选手在他们进行这项研究的 2005 年仍然健在——尽管这将使他们的年龄达到 138 岁！如果这 3 名选手死于 2005 年，那么 22 名选手的平均死亡年龄将提高到 90 岁。由于作者假定这三名选手在 2005 年依然健在，因此他们将获得更长的寿命。

所以，拥有已知死亡日期的选手平均比赛·扬早死 5 年，但如果错误地添加由于没有已知死亡时间而被认为在 138 岁时仍然非常硬朗的 5 名选手，那么赛·扬的同龄人似乎具有比他多 2 年的平均寿命。

艾贝尔和克鲁格犯下的另一个错误是，他们不是计算每位选手的死亡年龄，而是计算他们在入选名人堂以后活了多少年。这是有问题的，最明显的原因在于寿命预期取决于一个人的年龄。艾贝尔和克鲁格的程序没有对 87 岁入选名人堂的埃尔默·弗利克（Elmer Flick）和 47 岁入选名人堂的小卡尔·瑞普肯（Cal Ripken Jr.）进行区分。艾贝尔和克鲁格计算的选手入选名人堂以后在世的平均时间是一个没有用的统计量。相比之下，死亡年龄是一个更好的数字。

人们使用拥有已知死亡日期的选手的死亡年龄重新进行了分析。截至 2010 年，棒球名人堂有 292 人，其中 164 人在入选名人堂时仍然健在。在这 164 个名人堂成员中，62 人在 2010 年仍然健在，102 人已经故去。对于每个死去的名人堂成员，人们找到了出生在同一年并在这个成员入选名人堂时仍然健在的所有死去的选手。

同艾贝尔和克鲁格得到的结论相反，入选名人堂的选手比同龄人平均多活大约 1 年。不过，这种差异还够不上统计显著性的门槛。入

选名人堂的选手与同龄人的预期寿命差异不具有统计说服力。

×如何轻松识破一本正经地胡说八道

在对可发表理论的热切寻找过程中（不管这种理论多么不合理），人们很容易对数据进行微调，以便为理论提供更多支持。而且，如果统计检验给出人们所期待的答案，那么人们自然不愿意进行仔细的检查。如果研究人员在数据中搜寻某种模式，然后对不符合这种模式的数据进行篡改和删减，以便得到具有统计显著性的结果，那么你应当对这种研究保持警惕。如果统计结论看上去是不可信的，应对数据和检验进行检查。即使是专业人员也会犯错误。

① 小熊队刚刚在 2016 年打破了这个诅咒。——译者注

第 15 章
特异功能真的存在吗?

1972 年 6 月，5 个人闯进了华盛顿特区水门酒店民主党全国委员会总部。他们显然是想窃取民主党竞选计划的复印件，并在电话机和办公室的其他地方安装窃听设备。

水门酒店的一名保安注意到，一些门闩被黏上了胶带，以防止房门上锁。他叫了警察，他们抓到并逮捕了五名窃贼。所有人都相信，除了共和党人，没有人会闯进民主党委员会总部。如果你不是共和党人，那么那里的东西恐怕对你没有任何价值。此次盗窃行为是谁授权的？与此有关的共和党指挥系统中的最高领导人是谁？理查德·尼克松总统（President Richard Nixon）知道此事吗?

尼克松的新闻秘书轻蔑地将此次事件称为"三流盗窃未遂"。尼克松向全国人民表示："我可以断言……白宫和这届政府目前雇用的所有人员都没有参与这个极为古怪的事件。"尼克松由于肮脏而有些古怪的竞选策略而被长期称为"狡猾的迪克"，许多人怀疑这是他的又一个鬼把戏。

水门盗窃案演变成了水门丑闻。越来越多的证据表明，白宫参与了这项计划。更糟糕的是，尼克松也与这件事有关，他欺骗了全国人民。美国人往往会原谅那些承认错误的政客和名人，同时惩罚那些说谎的人。事实证明，有一名窃贼是共和党安全助手，尼克松重新竞选

总统的几万美元捐款存放在这些窃贼的银行账户里。这并不是三流盗窃。

美国参议院特别委员会召开了在全国范围内进行电视直播的水门听证会，时间一天天过去，罪证变得越来越多。最具戏剧性的事情完全出乎了人们的意料。尼克松曾在总统办公室、内阁会议室以及他的私人办公室里安装麦克风，以秘密记录人们的谈话内容。这些记录将彻底揭示尼克松是否参与最初的计划以及随后对窃盗案的掩饰。最高法院强迫尼克松移交这些磁带。虽然其中有 18 分钟 30 秒的内容被可疑地擦除，但还是有足够多的证据证明尼克松曾参与阻止联邦调查局的调查以及向窃贼支付封口费的计划。

众议院司法委员会投票建议以阻碍司法、滥用权力和藐视国会的名义弹劾总统。尼克松失去了所在政党的支持，他的出局显然已经不可避免了。所有人都认识到了这一点，除了一个之前默默无闻的国会议员，这位议员看不到自己不想看到的事情。

来自印第安纳州的代表厄尔·兰格雷布（Earl Landgrebe）的名声在很大程度上来自他在整个水门听证会过程中对理查德·尼克松总统的坚定支持。他曾告诉记者："不要用事实迷惑我；我不听。"他还说过："我将支持我的总统，即使我和他将被带出这座大楼并被射杀。"第二天，尼克松辞职。兰格雷布成了《纽约时报》杂志 1974 年"十大愚蠢"国会议员之一。

兰格雷布当然不是唯一具有封闭心态的政客。封闭心态也不是政客的专利。据说，即使是本应具有冷静头脑的科学家也像艺术家一样深爱着自己的模型。我们都有自己的理论。为什么股市下跌。为什么一家公司取得成功。为什么某人得到提拔。而且，我们都想强调与我们的理论相符的数据，忽略与之相矛盾的数据，或者将那些与我们的理论相冲突的数据错误地解读成与之相符的数据。不过，当那些本应具有客观态度的科学家像其他普通人一样无法看到局外人所能看到的事情时，人们会产生更加强烈的失望和幻灭感。

超感知觉

超感知觉（ESP）是在不使用视觉、听觉、味觉、嗅觉和触觉这五种物理感官的情况下获取信息的能力。超感知觉包括心灵感应（阅读另一个人的思想）和千里眼（识别看不见的物体）。这些通灵能力通常是即时的，但也可能涉及预先感知——获得还没有发生的某件事情的信息。

我们可以使用心灵主义者在魔术表演中使用的把戏来展示这些通灵能力：

- 心灵感应：志愿者从一副洗好的纸牌中选择一张牌，看清这张牌，然后将其放回原处。心灵主义者说，他能阅读志愿者的思想。然后，他指出志愿者选择的那张牌，以证明这一点。
- 千里眼：观众席成员写下与自己有关的便条（也许是死去的亲人的名字），然后将其放进帽子里。心灵主义者说，他能看到眼睛看不到的东西。接着，他把纸条一个一个从帽子里取出来，紧紧攥在手中，然后识别出纸条上的内容。
- 预先感知。心灵主义者说，他能预测出志愿者在四种花色中选择哪一种花色。在志愿者选择一种花色（用眼睛看正面朝上的牌或者将背面朝上的牌翻过来）以后，心灵主义者打开之前封好的、放在明处的信封，拿出里面的纸条，上面所写的花色与志愿者的选择相符。

这些都是魔术师所使用的标准戏法。魔术师并没有特殊能力，但是却能提供令人信服的娱乐表演。

这些通灵能力真的存在吗？或者，它们只是巧妙的戏法？除了魔术师，没有人能够以 100% 的准确度证明这些能力。不过，一些严肃的研究人员相信，普通人能够一致地做出超越随机猜测的表现。

最著名的超感知觉研究员是 J. B.莱因。莱因本来是芝加哥大学植物学博士研究生。在听了亚瑟·柯南·道尔爵士发表的一场关于唯灵论的演讲以后，他突然改变了自己的事业方向。道尔的名声来自他笔下的夏洛克·福尔摩斯的故事。福尔摩斯是小说中的侦探，以其对犯罪现场的细致研究和非凡的逻辑推理能力著称。

道尔与他所创造的福尔摩斯这一角色完全相反。他痴迷于超自然现象（根据定义，这种现象没有合理的解释）。他的信念极为强烈，这使他对可能威胁自身信仰的所有证据视而不见。

英格兰柯廷利镇的两个小女孩 1917 年到 1920 年拍摄了五张仙女照片。图 15.1 显示了其中的第一张照片。这也许是最令人尴尬的一个例子。



图 15.1　仙女是真实的（圣诞老人也是）

这些仙女看上去是二维的，她们的亮度很奇怪，她们的翅膀是静止的，整张照片看上去缺乏可信度。不过，道尔仍然对仙女照片的真实性深信不疑，他还为此写了一本书，叫做《仙女的到来》。多年以后，在道尔去世很久以后，其中一个女孩承认了显而易见的事实。仙女的形象是从杂志上剪下来并用帽针固定在她们身边的。

道尔还坚定地相信，活人可以和死人交流。现代招魂术始于福克斯三姐妹，她们是生活在纽约农村的青少年。在降灵会中，她们在黑暗的屋子里和单纯的客户坐在一张桌子周围，死者会用叩击声与活人交流。当其中一个姐妹最终承认这是圈套时（神秘的叩击声是她活动脚趾关节形成的！），道尔拒绝相信她："她在这方面的任何说辞都不会在任何程度上改变我的观点。"

魔术师哈里·霍迪尼（Harry Houdini）试图使道尔认识到他受到了骗子的利用。不过，道尔并不相信他。当霍迪尼向道尔展示冒牌巫师所惯用的把戏时，道尔认为霍迪尼在掩饰自己的超能力。道尔甚至认为霍迪尼在著名的逃脱魔术中失去了身体形态。我们每个人都会抗拒威胁自身信仰的证据，但道尔的天真和自我欺骗达到了令人震惊的程度。

我们可能认为道尔对仙女照片的盲目相信和他对福克斯姐妹所承认的事实的坚决否认会毁掉他的名声，尤其是他在科学家中间的名声。事实并非如此。在听到道尔关于唯灵论的芝加哥演讲以后，J. B.莱因非常兴奋。他后来回忆道："'与死者交流的'这种可能性是我多年来有过的最愉快的想法。"

在获得植物学博士学位后不久，莱因进了哈佛大学，跟随威廉·麦克杜格尔教授（Professor William McDougall）学习心理学。麦克杜格尔刚刚成为支持通灵现象研究的两家机构的主席。在波士顿期间，莱因和他的妻子参加了灵媒米娜·（"玛杰里"）·克兰登举办的一场降灵会。米娜当时非常有名，但是现在已经变得声名狼藉。莱因一家人产生了怀疑（为什么降灵会总是在黑暗中进行呢？）。事后，他们发表

了一篇详细的揭露性文章，解释了他们为什么认为“整个事件是一场低劣而无耻的骗局”。

亚瑟·柯南·道尔被激怒了。他给《波士顿先驱报》写了一封信，谴责莱因一家人“巨大的无礼”，并且花钱在波士顿报纸上刊登广告，上面只写着“J. B.莱因是个十足的蠢货”。在其他时候，道尔完全不愿意考虑可能危及自身信仰的证据。此次事件的另一个有趣之处是，它显示了莱因在职业生涯早期对通灵现象持有的开放心态。莱因真诚地希望调查这些现象的真实性。不过，他的判断很快受到了“相信的愿望”的蒙蔽。

在哈佛待了一年以后，莱因在 1927 年跟随麦克杜格尔来到了杜克大学。在这里，麦克杜格尔成了心理学系主任，并且支持莱因成立通灵学实验室。莱因在这个实验室对超感知觉进行了几十年的研究。

他最著名的实验涉及我们在第二章提到的齐纳纸牌。一副齐纳纸牌包含 25 张牌，分为 5 种图案：圆圈，十字，波浪线，方块和星星。每种图案 5 张牌。在典型的实验中，一个人（发送者）洗牌，然后依次将 25 张牌翻过来，盯着每张牌看一会儿，然后记录这张牌的符号。另一个人（接收者）努力识别每张牌。有时，接收者在测试过程中被告知每次猜测是否正确；其他时候，发送者直到整副牌的猜测结束后才会揭晓正确答案。不管是哪一种方式，答案都会得到记录，测试结束后还会有统计分析。

莱因 1934 年的《超感官知觉》一书描述了他在杜克大学的实验中得到的充满希望的结果。《纽约时报》科学编辑夸张地赞美道：“他的工作所具有的价值是毫无疑问的……由于他的原创性、他那严谨的客观性以及他对科学方法的严格遵守，他为自己赢得了名声。”全国各地的报纸和杂志涌现出了许多关于超感知觉的文章。莱因出名了。许多普通市民购买 10 美分一副的齐纳纸牌，然后在家中进行超感知觉实验。超感知觉的热潮很快蔓延开来。

欧文·朗缪尔（Irving Langmuir）是一位杰出的化学家和物理学家，他在 1932 年获得了诺贝尔化学奖。朗缪尔非常尊重科学，对没有

经过严格检验的令人难以置信的说法抱有深深的怀疑态度。在莱因的《超感官知觉》一书出版的时候，朗缪尔拜访了莱因的杜克实验室，他想亲自看一看到底发生了什么。几年以后，1953 年，朗缪尔发表了一篇题为“病态科学”的演讲。“病态科学”一词是他发明的，用于描述“名不副实的科学”。在演讲中，朗缪尔回忆了他访问莱因实验室的情景。朗缪尔吃惊地发现，莱因正在丢弃一部分数据，将几万个装有齐纳纸牌的信封放进文件柜里，因为他感觉一些参与者故意给出了错误的答案！如果你对某件事的信念达到了一定程度，你总能找到借口。

中立的研究人员应当分析所有数据，包括正确的预测和错误的预测，包括正面结果和负面结果。例如，如果接收者在使用 25 张齐纳纸牌的测试中只能进行随机猜测，我们认为接收者平均可以做出 5 次正确猜测。（实际情况稍微复杂一些，因为接收者知道一副纸牌只有 5 张圆圈、5 张十字……不过，5 次正确猜测的预期基本是正确的。）

假设一个人做出 7 次正确预测，另一个人做出 3 次正确预测。公正的观测者会说，他没有看到不同寻常的现象，因为两个人的平均结果是 5 次正确预测。不过，寻找超感知觉证据的人可能会说，第一个人做出了超过 5 次的正确预测，因此证明了超感知觉；第二个人通过做出不正确的预测成功羞辱了莱因，因此他也证明了超感知觉。

这就是莱因的观点。他强烈相信超感知觉，因此他觉得将正面和负面结果分开并将每一种结果看作超感知觉证据的做法是完全合理的。正面结果说明人们拥有超感知觉。负面结果说明人们拥有超感知觉，而且他们正在利用自己的超感知觉羞辱他。

朗缪尔走了，他相信莱因是自我欺骗的诚实受害者。莱因的实验没有证明超感知觉；相反，它证明了一些出于好心的研究人员无法意识到他们所制造的恶作剧。

朗缪尔还说，当他告诉莱因他认为莱因的超感知觉研究是病态科学时，莱因回答道：“我希望你发表这种观点。我很愿意让你发表这种观点。那将激发许多人的兴趣……我将拥有更多研究生。我们应当拥

有更多研究生。这件事情非常重要，因此我们应当让更多的人意识到它的重要性。这个系应当成为学校里的一个大系。”莱因无意中暗示了自己的身份。他不再是试图确定超感知觉真实性的公正的研究人员，而是追求名誉和资助的派系拥护者。

超感知觉实验还是选择性报告的一个经典案例。即使超感知觉不存在，平均每 20 个受试者之中也会有 1 个人做出足够多的正确猜测，得到具有统计显著性的结果。在数千人的反复测试中，总会出现一些令人难以置信的结果——只有这些值得注意的结果会得到报告。如果我们知道一共进行了一千次测试，那么其中一次测试得到具有千分之一可能性的结果也就不足为奇了。

在 1953 年的演讲中，朗缪尔提到一个个人案例，这个案例涉及他的侄子戴维·朗缪尔（David Langmuir）：

> 戴维和一群年轻人想要检验莱因的工作，因此他们找了一些纸牌，花了许多个夜晚共同猜测纸牌的图案，他们的猜中次数远远高于 5。他们感到非常激动，不断进行实验，而且准备向莱因写信报告这件事。不过，他们稍微拖延了一段时间。他们的猜中次数开始一点一点地持续下降。经过许多天以后，他们的总体平均结果下降到了 5 次，因此他们并没有给莱因写信。如果莱因收到了这条信息，也就是这群很有信誉的人在许多次实验中得到了 8、9 或 10 的总体结果，那么他一定会把这件事写进他的书里。有多少对此感兴趣的人会为你反馈这样的信息呢？你如何对书中发表的这类信息进行衡量呢？

这种选择性报告也叫“出版效应”，因为具有统计显著性的结果会被写进期刊和图书，而不显著的结果则不会得到报告。

对抗数据挖掘和选择性报告的第一种良方是常识。许多人相信超感知觉，但是其他人并不相信。也许，最安全的说法是，如果超感知

觉的确存在，那么它似乎并没有太大的现实意义。例如，没有证据表明人们能够利用超感知觉使拉斯韦加斯、大西洋城或者蒙特卡洛的赌场破产。

第二种良方是新数据。得到较高分数的受试者应当进行重新测试。莱茵报告说，当他对得到高分的人进行重新测试时，他们的分数几乎一定会下降。由于这种现象经常发生，因此莱因给它起了一个名字，叫做“下降效应”。他的解释是：“他们显然感到了疲惫和厌倦。”另一种解释是，最初的高分仅仅是对幸运猜测的选择性报告而已。

巨大的测试人数并不是对实验结果的统计显著性进行评估的唯一困难。莱因还考察了受试者在一项测试中表现出色的一个片断，这也是一种数据挖掘。5% 受试者做出的优异表现以及一个人在一项测试的 5% 结果中做出的优异表现并不使人感到吃惊。

莱因还考察了“前向位移”和“后向位移”，也就是一个人的选择与当前的纸牌不匹配，但是与后一张纸牌、前一张纸牌、后面第二张纸牌或者前面第二张纸牌相匹配。这种多重潜在匹配提高了找到巧合模式的可能性。莱因还认为一个人在一项测试的某些部分得到高分、在另一些部分得到低分（“反向超感知觉”）的现象是不同寻常的。莱因解释说，一位受试者可能在测试开始时位于均值一侧，并在测试结束时偏向均值另一侧同样远的距离；或者，他可能在测试中间处于较低水平，在两边处于较高水平。两种偏差趋势可能会相互抵销，使整个测试的平均水平接近“概率”。

他对一次测试结果的描述是：

> 当其中一位发送者看牌时，前向位移和后向位移同时存在，当另一个人充当发送者时，只存在前向位移；目标的位移移动一张还是两张纸牌取决于测试的速度。

即使猜测是随机的，面对如此众多的受试者和如此众多和可能性，你也很容易找到模式。

莱因的研究给我们带来的最主要的教训是，我们应当对基于搜刮数据的研究保持警惕。通过各个角度研究数据的人一定会发现某种现象。这种现象只能证明他们通过各个角度对数据进行了研究。

莱因的超感知觉研究令人兴奋，但他无可厚非的热情却使数据挖掘和选择性报告的问题变得非常严重，这毁掉了研究的可信性。如果他能多一些对统计检验的正确使用，少一些滥用，超感知觉研究就会变得更具说服力，或者不会产生那么大的争议。1996 年，自称是怀疑主义者的魔术师詹姆斯·兰迪（James Randi）悬赏 100 万美元；只要能够“在适当观测条件下证明任何超常、超自然或神秘的能力或事件的证据”，任何人都可以获得这笔奖金。目前，这笔奖金仍然无人认领。

霍迪尼的挑战

在职业生涯早期，哈里·霍迪尼伪装成了一个灵媒，自称能够传达来自灵界的消息。有的时候，霍迪尼让人把他绑在“幽灵小屋”中的一把椅子上，他的脚边放着铃铛和铃鼓。在他装作入定以后，一个帘子会垂下来。接着，他的器具会发出巨大的声响，然后从观众的头顶飞过。当帘子打开时，人们发现霍迪尼仍然被牢牢地绑在椅子上。这当中的把戏是，如果使用合适的捆绑方式，你很容易根据需要把双手从绳套里滑出来或者滑进去，这样你就可以摇动铃铛和铃鼓并把它们扔到小屋外面了。（我之所以知道这一点，是因为一个灵媒曾经把我从观众席叫上台，让我和她待在一个幽灵小屋里。当我们进入小屋的时候，她让我配合她的行动。我配合了。）这个把戏的简单程度和大胆程度令人感到好笑。不过，如果合适地呈现在愿意相信灵魂的观众面前，这个把戏是非常令人震撼的。

有的时候，霍迪尼会进入入定状态，然后转达死者的消息。这当中的把戏是，霍迪尼事先和当地居民谈话，阅读以前的新闻，在当地的墓地抄写姓名和日期，以便整理出这些消息。据说，在一场极为真实的表演中，当霍迪尼拼出某个被割喉的死者的名字时，人们吓得四散奔逃。

后来，霍迪尼心爱的母亲去世了，他拼命想要和她取得联系。他找了许多巫师，但他每次都很失望。他轻而易举地戳穿了这些巫师用来欺骗小白的伎俩。而且，没有一个巫师能够猜出他的母亲去世之前说出的最后一个词语（“原谅”）。他开始迷上了揭发灵媒的工作，指出他们是“捕食丧亲之人的秃鹫”。经过乔装打扮，霍迪尼带着一位记者兼警察参加降灵会。当他弄清对方的把戏时，他就会去掉伪装，大声宣布：“我是霍迪尼，你是骗子！”

他还将招魂术用到了舞台表演之中，以揭发最著名的灵媒使用的伎俩。他悬赏 1 万美元，请人们演示无法被他复制的超自然现象。没有人认领这笔奖金。

作为最后的挑战，霍迪尼和妻子贝丝（Bess）约定了一个秘密消息，两个人之中最先去世的人需要发送这个消息。消息的第一部分是“罗莎贝尔”。这是他妻子的昵称，而且被刻在她的结婚戒指内面上。消息的第二部分基于他们共同表演时使用的密码。这种密码用十个词语表示数字：

1. 祈祷
2. 回答
3. 说
4. 现在
5. 告诉
6. 请

7. 讲

8. 快

9. 看

10. 快点（也可以表示 0）

例如，贝丝可能会走进观众席，从一个志愿者手中接过一枚硬币，然后问哈里："请告诉我日期的最后两位数字。"听到"请"和"告诉"两个词语，哈里知道正确答案是 65。在其他特技表演中，每个数字对应一个字母；例如，"回答"表示字母表中的第二个字母（B），"现在"表示第四个字母（D）。对于字母表后面的字母，他们可以将词语组合在一起，并用停顿表示字母的结束。因此，一个包含"回答"（2）和"祈祷"（1）的短语表示字母表中的第 21 个字母 U。

哈里（或者贝丝）需要在死后发送的加密消息是"回答"（B）"告诉"（E）"祈祷回答"（L）"看"（I）"告诉"（E）"回答回答"（V）"告诉"（E）解码后的完整消息是"罗莎贝尔，请相信（believe）。"

在霍迪尼 1926 年去世以后，贝丝悬赏 1 万美元，以奖励能够接收霍迪尼发来的 10 个词语（"罗莎贝尔"加上 9 个密码词语）的人。她举行了数百场降灵会，包括一年一度的万圣节降灵会，但是都没有取得成功。1929 年，贝丝宣布，一位名叫亚瑟·福特（Authur Ford）的巫师破解了霍迪尼的密码。不过，两个人后来承认，这是他们捏造的谎言。贝丝说，这是精心制造的骗局（也许是因为她和福特的恋情，或者因为她希望得到宣传）；福特说，在他的欺骗下，贝丝泄漏了这个秘密。不管真相如何，福特并没有获得奖金。

1936 年，另一场失败的万圣节降灵会在好莱坞具有传奇色彩的纽约人酒店顶楼举行，并且在全世界范围内得到了无线电广播。随后，贝丝吹灭了一直燃烧在霍迪尼照片旁边的蜡烛。她后来表示，"对于任何人来说，十年的等待已经足够长了。"

一个家庭对超自然现象的着迷

1872 年，弗朗西斯·高尔顿爵士（Sir Francis Galton）报告说："一位著名的权威人物最近发布了一项挑战，请人们用真实的实验来检验祈祷的效力。"为了回应这项挑战，高尔顿提出，他可以比较由于骨折和截肢而住院的两个群体的恢复情况："一组包含明显非常虔诚、能够受到其他人好心帮助的个体，另一组包含明显很冷淡、被人忽略的个体。"

高尔顿并没有开展这项研究，但他报告了著名英国教士和英国王室的寿命，这些人每个星期日都会接收到来自全英格兰各个教堂的公众祈祷："祝愿他/她健康长寿。"经过计算，高尔顿发现，教士的预期寿命（66.4 年）略低于律师（66.5 年）和医生（67.0 年）。王室成员是英国接受祈祷最多的人，他们的预期寿命只有 64.0 年，明显低于其他贵族（67.3 年）。如果祈祷能够产生影响的话，那么它似乎弊大于利。

高尔顿总结道："在文明世界，大量诚实的信念已经让步于确凿事实的无情要求；在我看来，从我所考虑的意义上说，对于祈祷效力的所有信念显然也必须做出让步。"如果翻译成正常的英语，这句话的意思是，祈祷不会延长寿命。

不过，高尔顿的比较并不是真正具有科学性的比较。他提议观测能够得到他人祈祷和被人忽略的住院病人，但是这两个群体之间可能存在系统性差异。也许，"很冷淡、被人忽略的人"往往更加贫穷，年纪更大，或者不太注意卫生。而且，高尔顿提出的比较并不是双盲的（即患者和医生都不知道谁在接受祈祷）。如果医生知道谁在接受祈祷，这可能会影响他们对病人恢复情况的客观评估。

高尔顿对教士和英国王室与律师、医生和贵族的比较有一个问题，那就是人们并不是随机分配到这些群体之中的。他们进入这些群体的原因包括血统和选择。也许，王室基因存在某种（好的或坏的）

独特之处。也许，选择成为律师或医生的人与教士之间存在系统性差异（也许他们更加积极或富有）。也许，医生更加关心自己的健康，更加了解合适的治疗方法。高尔顿提出的研究和他所进行的研究都存在严重缺陷。

×如何轻松识破一本正经地胡说八道

不同寻常的说法需要不同寻常的证据。笃信者则不需要这么高的要求。

如果亚瑟·柯南·道尔能够拒绝相信骗子承认自己假装与死者交流的说法，那么任何人都有可能这样做。如果 J. B.莱因能够认为志愿者没能正确猜出卡片的现象是志愿者正确猜出卡片的证据，那么任何人都有可能这样做。对于犯错的可能性持有包容的心态是一件很难做到的事情。不过，我们应该努力做到这一点。

第 16 章
彩票是一种智商税

数据仅仅是数据而已。即使我们看到清晰而明确的模式，要想相信这种模式不仅仅是巧合，我们仍然需要一种合乎逻辑的理由。有时，我们很容易找到某种解释，而且完全有理由相信这种模式将会持续下去。由于感恩节的原因，几十年来，美国的火鸡消费量往往会在 11 月出现激增，这种现象很可能还会持续多年。由于寒冷的天气，美国北方的建筑活动在冬季会变少，这种现象很可能还会持续下去。

其他时候，我们所拥有的仅仅是没有理由的数据——这是很危险的。有人曾对卢旺达手机通话进行善意的监控，发现人们在村庄外部活动变少的几个星期以后往往会爆发霍乱。也许，当最初的症状出现时，人们刚好待在家里？谁知道呢？我们没有合乎逻辑的解释，但是谁在乎这一点呢？重要的是，这个项目无意中找到了一种预测霍乱爆发的出人意料的方法。不是吗?

不见得。实际上，人们之所以停止离开村庄，是因为突然爆发的洪水冲毁了道路，而这些洪水也提高了霍乱的风险。即使不像“老大哥”那样监督公民的一举一动，卢旺达政府也很容易观测骤发洪水，而且手机数据无法预测出并非由冲毁道路的洪水导致的霍乱爆发。

类似地，如果我们仅仅根据过去的趋势推测未来，而不去考虑这种趋势是否有意义，那么我们的结论可能会与众所周知的真相相去甚

远。如果我们对股票价格和彩票中奖数字进行仔细检查，寻找跑赢大盘和中彩票的荒谬办法，我们几乎一定会得到更加糟糕的结果。

缺乏理论的数据可能会引发股市投机泡沫，或者在没有泡沫时创造出存在泡沫的幻象。如何分辨真实泡沫和虚假警报之间的区别呢？你知道答案的：我们需要一种理论。数据本身是不够的。

目之所及

亚伯拉罕·林肯在第二篇国情咨文中计算了从 1790 年到 1860 年每个十年人口普查之间的美国人口增长率。平均的十年增长率是 34.60%。林肯注意到，这 7 个十年的增长率与 34.60% 的偏差不超过 2%，“这说明我们的增长法则极为固定，因而极为可靠。”根据每十年 34.60% 这一数字，林肯预计，70 年以后，即 1930 年，美国人口将达到 251680914 人。注意，除了轻率的外推，他的预测还具有毫无理由的精确性。如果我们无法统计今天的准确人口，我们怎么能预测 70 年以后的准确人口呢？

实际上，美国 1930 年的人口是 1.23 亿，不到林肯预测值的一半。1938 年的一个总统委员会犯了相反的错误，认为未来的人口永远不会超过 1.40 亿。仅仅 12 年后，1950 年，美国人口已经达到了 1.52 亿。

如果我们对于某种历史趋势没有合乎逻辑的解释，但是仍然认为它会持续，这说明我们正在进行轻率的外推，这种外推完全有可能产生令人尴尬的错误。有时，轻率的外推仅仅是为了逗人一笑。一项研究考察了过去 350 年的英国演讲家，发现平均句子长度从弗朗西斯·培根的每句 72.2 词下降到了温斯顿·丘吉尔（Winston Churchill）的每句 24.2 词。按照这个速度，每句单词数量将在一百年后达到零点，然后变成负值。

我们都将为 IBM 工作

1924 年，计算制表记录公司抛弃了笨拙的名字，换上了一个更有进取心的名字——国际商业机器（IBM）。此后，公司成了持续的一流增长股。到 1978 年，公司业绩已经在超过 50 年的时间里实现了每年约 16% 的增长（扣除通胀因素）。一些财务分析师根据过去的表现而不是未来的前景认为购买 IBM 的股票永远不会是错误的决定。当时有一种流行的说法："没有哪个采购经理因为购买 IBM 的计算机而被解雇，没有哪个组合基金经理因为购买 IBM 的股票而被解雇。"

图 16.1 显示了 1978 年一份对于 IBM 未来十年每股收益的预测，它所依据的是过去十年的数据。一条平滑的曲线与这些数据拟合得很好，而且暗示 IBM 公司 1988 年的每股收益将达到 18.50 美元，是 1978 年收益的三倍。根据这些外推，许多股票分析师建议购买 IBM 的股票，认为它的价格将在未来十年增长两倍。

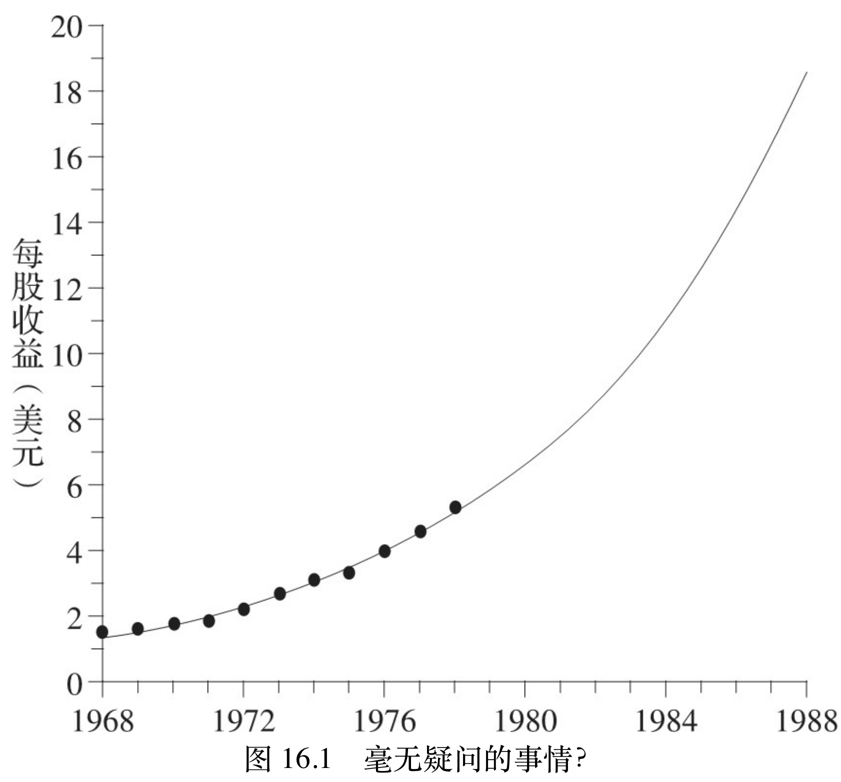


图 16.1　毫无疑问的事情?

和所有历史图表一样，图 16.1 仅仅具有描述性质。在我们根据过去的趋势做出充满信心的预测之前，我们应当超越这些数字，思考过往趋势的内在原因未来能否持续。

如果分析师考虑到这一点，他们可能会意识到，一些令人信服的理由可以说明 IBM 的高速增长无法永远持续下去。IBM 最初很小，并且随着计算机对整个经济体的渗透而迅速增长。到 1978 年，IBM 已经成了一家很大的公司，持续增长的空间已经不多了。同小公司相比，大公司每年增长 16% 的难度要大得多。

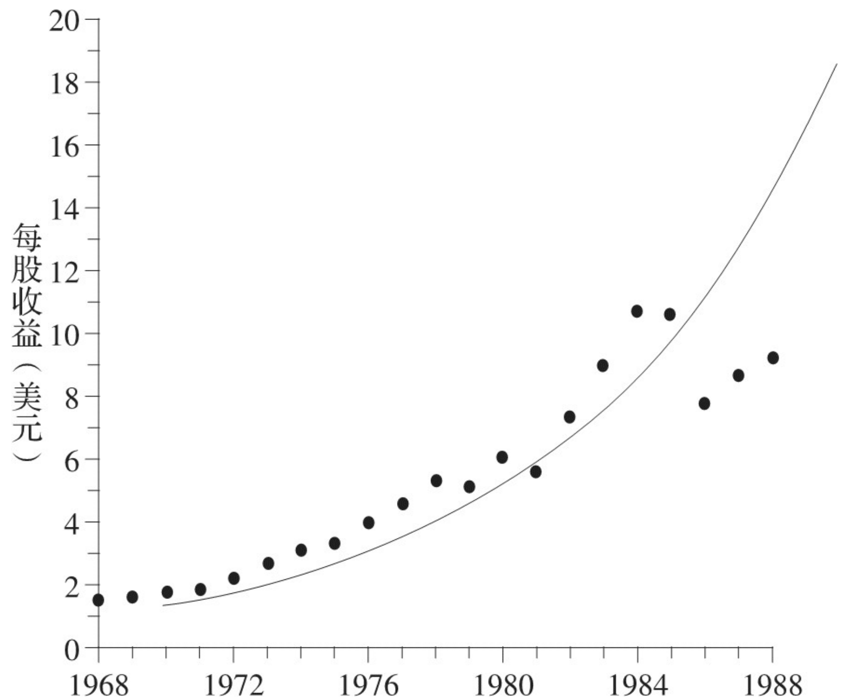


图 16.2 哎哟！

如果 IBM 继续以每年 16% 的速度增长，整个美国的经济继续以 3% 的长期速度增长，到 2003 年，美国的一半产出将是 IBM 的产品，到 2008 年，美国的全部产出将由 IBM 提供！在这种异想天开的推理中，我们需要放弃某个条件——IBM 的增长率需要下降到 3%，或者经济增长率需要上升到 16%。整个经济体维持 16% 的增长速度是非常

令人难以置信的，因为经济增长受到劳动力和生产力的限制，而且我们很难看出这两个因素的增长能够支撑起 16% 的经济增长率。

表 16.2 说明，IBM 公司 16% 的增长率并没有持续。事实证明，对于 1968～1978 年收益趋势的简单外推是一种轻率的外推，具有盲目的乐观性。IBM 公司 1988 年的每股收益不是 18.50 美元，而是这个数字的一半。随后的年份与之基本持平。IBM 不可能永远以 16% 的速度增长。那些在 20 世纪 70 年代购买 IBM 的股票、相信 IBM 引人注目的增长速度永远不会停止的投资者在失望中明白了一个道理：你很少能够通过后视镜看到未来。

股票是一种不错的投资

制作于 2008 年 12 月的图 16.3 说明，股票是一种极好的投资。从 2004 年 10 月到 2007 年 10 月，股票价格增长了 38%。而且，这种增长平滑而稳定——根本不像关于股市波动的新闻故事讲述的那样可怕。

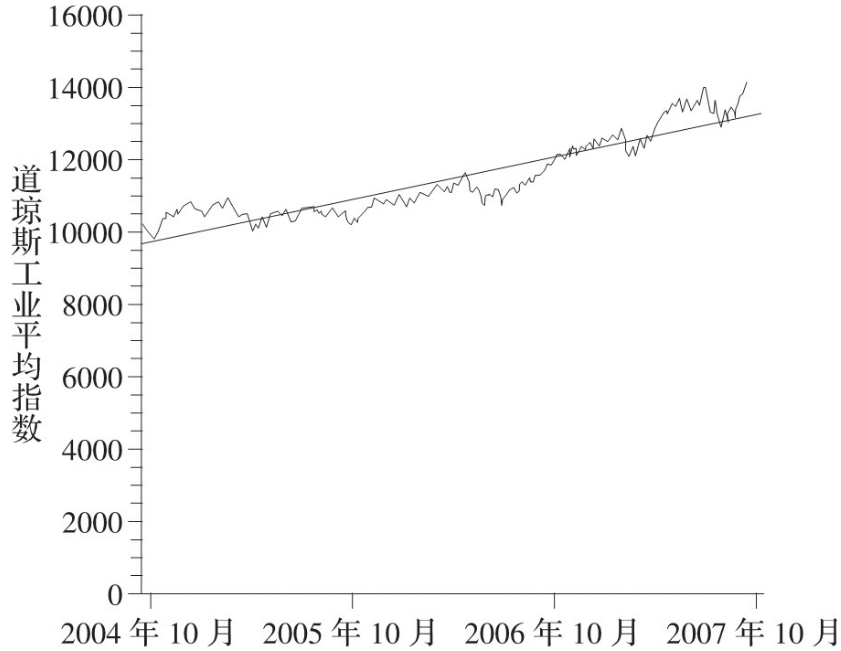


图 16.3　股票是一种不错的投资

既然 2008 年的数据已经出来了，为什么图 16.3 终止于 2007 年 10 月呢？因为后面的数据（见图 16.4）讲述了作者不想讲述的故事——从 2007 年 10 月到 2008 年 12 月，股票价格下跌了 46%。图 16.4 说明股票是一种糟糕的投资。



图 16.4　股票是一种糟糕的投资

对于图像起止时间的仔细选择可以创造出更加完整的图像中不存在的趋势幻象。哪些数据更加真实，是 2004 年 10 月到 2007 年 10 月的数据，还是 2007 年 10 月到 2008 年 12 月的数据？两个时间段都很短暂，无法对变化无常的股票价格做出真实的总结。正如一位诚实的预测者所说："股票价格将会上涨。股票价格将会下跌。也可以把这个顺序颠倒过来。"

要想获得均衡的观点，需要使用长期视角。过去一百年，股票价格平均每年增长大约 4%。

会有人天真到根据短短几年的数据对股票价格进行外推吗？投资者一直在做这样的事情，他们有时依据的时间更短。他们希望最近的反弹是大量幸福利润的前奏，或者担心最近的回挫是崩溃的开始。2007 年 10 月，我所认识的一位财务规划师的许多客户开始借钱，希望能在股市上获得两位数的回报。15 个月后，他们卖掉了一切，因为他们不想在股市上损失另一个 50%。

每当有人出示具有奇特起止时间点的数据时，我们都应当产生怀疑。如果 2009 年制作的图表使用了 2004 年 10 月到 2007 年 1 月的数据，我们应当产生疑问：为什么他们不使用 2004 年 10 月之前或者 2007 年 1 月之后的数据？为什么他们的数据起始于 10 月，终止于 1 月？如果起止点看上去是对数据进行仔细研究之后做出的特殊选择，那么这些选择很可能是为了歪曲历史记录。它们也许拥有合乎逻辑的完美解释，但我们应该坚持要求听到这种解释。

跑赢大盘（或者说为什么我喜爱包装工队）

我们寻找模式和秩序的内在愿望与我们通过购买正确的股票或正确的彩票发财的如意算盘很好地结合在了一起。问题是——持续存在的问题是——即使股市基本是随机的，彩票完全是随机的，我们也总是能够在随机数字中找到一些模式。只要我们寻找模式，我们就能找到它。

在 1983 年 1 月的超级碗星期日，《洛杉矶时报》的商业版和体育版都刊登了关于“超级碗预测股市”的文章。这种理论认为，如果国家橄榄球联合会（NFC）或者目前位于美国橄榄球联合会（AFC）中的前国家橄榄球联盟（NFL）球队赢得超级碗，股市将会上涨；否则，股市将会下跌。对股市来说，绿湾包装工队的胜利是好消息，纽约喷气机队的胜利则是坏消息。

在前 16 场超级碗中，这种理论成功了 15 次。《洛杉矶时报》引用了一位股票经纪人的话：“市场观察员将盯住电视屏幕……你很难忽视一个准确率超过 94% 的标普指标。”华盛顿红人队（NFC 球队）取得了胜利，股市出现了上涨。第二年，“超级碗指标”再次出现在了新闻中，这一次的宣传势头比以往更加强烈。截至 1997 年，这种超级碗

理论在 31 次比赛中实现了惊人的 28 次成功。随后，它在接下来的 14 年里失败了 8 次。

股市与橄榄球比赛的结果没有任何关系。超级碗指标的准确率仅仅是一种可笑的巧合，因为股市经常上涨，而 NFC 也经常取得超级碗的胜利。为了使这种相关性更加明显，人们耍了一个花招，将 AFC 的匹兹堡钢人队算作 NFC 球队。他们给出的表面上的理由是，匹兹堡队曾经是一支 NFL 球队；实际上，这是因为匹兹堡队曾经多次在股市上涨时获得超级碗冠军。将匹兹堡队算作 NFC 球队的做法是对数据的歪曲，其目的仅仅是为了支持这种荒谬的理论。

1989 年，在“超级碗指标”这一成功案例的激励下，《洛杉矶时报》撰稿人对数据进行了更加严重的歪曲，发现了一些类似的巧合。他的“哟，阿德里安理论”认为，当洛基或兰博的电影发布时，股市将会上涨。

同年，《纽约时报》转换了预测方向。除了用超级碗预测股市，为什么不能用股市预测超级碗呢？为什么不呢？这并不比最初的超级碗指标更加可笑。《纽约时报》报道说，如果道琼斯工业平均指数从 11 月末到超级碗比赛时出现增长，那么城市名称按字母顺序排在第二位的橄榄球队通常能够赢得超级碗。（提示：你觉得他们为什么将起始日期选在 11 月末，而不是 1 月 1 日、比赛一个月前、比赛一年前或者其他任何合乎逻辑的日期？）

再后来，贝斯普克投资集团发现了《体育画报》泳装指标：如果《体育画报》年度泳装特刊上的封面模特是美国人，股市就会出现良好的增长势头，如果模特是外国人，股市就不会表现得太好。

自从超级碗指标被人发现以来，它的表现很平凡——这并不奇怪，因为它仅仅是巧合而已。真正令人吃惊的是，一些人并没有理解这个笑话。提出超级碗指标的人想通过一种幽默的方式说明相关性并不意味着因果关系。他没有想到的是，人们竟然接受了这种理论！

傻瓜四股

1996 年，加德纳兄弟写了一本非常流行的书，书名很有诱惑力，叫做《彩衣傻瓜投资指南：傻瓜带你战胜华尔街聪明人》。嘿，如果傻瓜可以跑赢大盘，那么我们所有人都能做到这一点。

加德纳兄弟将他们的建议称为“傻瓜四股”策略。他们说，从 1973 年到 1993 年，“傻瓜四股”策略的年均回报率是 25%，而且这种策略“未来将为它的粉丝带来 25% 的年回报率，就像过去那样。”

下面是投资致富的秘方：

1. 在年初计算道琼斯工业平均指数 30 只股票中每只股票的股息率。例如，2014 年 1 月 2 日，可口可乐股票的每股价格为 40.66 美元，年度股息为每股 1.22 美元。因此，可口可乐的股息率为 1.22 美元/40.66 美元=0.0300，即 3.00%。
2. 在 30 只道指股票中，确定股息率最高的 10 只股票。
3. 在这 10 只股票中，选择每股价格最低的 5 只股票。
4. 在这 5 只股票中，划掉价格最低的股票。
5. 向价格第二低的股票投入 40% 的财富。
6. 向其他三只股票各投入 20% 的财富。

正如戴夫·巴里（Dave Barry）所说，这不是我杜撰出来的。

有人感到奇怪吗？为什么这种策略复杂到了近乎令人困惑的程度？也许，这是一种数据挖掘？

步骤 1 和步骤 2 看上去是合理的。股息率较高的股票很有吸引力。“道指狗股”是一个历史悠久的投资策略，倾向于购买股息率最高的道指股票。这种合理的策略取得了一定的成功，它的基本思想是，

逆向投资者应当逆势而行，购买价格较低、股息率较高的冷门股票。超级投资者沃伦·巴菲特（Warren Buffett）曾经说过一句令人难忘的话："当其他人贪婪时，你应当感到恐惧；当其他人感到恐惧时，你应当贪婪。"

不过，除了这种借来的思想内核，"傻瓜四股"完全是一种数据挖掘。步骤 3 没有合乎逻辑的基础，因为每只股票的价格取决于公司公开发售的股票数量。如果公司将股票数量加倍，每股的价值将缩小一半。没有任何理由能够解释为什么股票发售数量较多（每股价格较低）的道指股票比股票发售数量较少（每股价格较高）的道指股票更加优秀。稍后，我们将在这一章看到，沃伦·巴菲特的伯克希尔·哈撒韦公司（不在道指之中）公开发售的股票数量非常少，因此价格达到了惊人的每股近 20 万美元。不过，它仍然是一项不错的投资。

步骤 4 呢？在选择价格最低的 5 只股票（似乎低价是一个有利因素）以后，为什么要划掉价格最低的股票呢？这到底是为什么呢？

步骤 5 和步骤 6 呢？为什么价格第二低的股票投资金额是其他三只股票的两倍呢？我们都知道答案。因为它在历史上是有效的。仅此而已。

在加德纳兄弟提出"傻瓜四股"策略后不久，两位持有怀疑态度的金融教授用加德纳兄弟挖掘过的数据之前的 1949 ~ 1972 年的数据对这种策略进行了检验。结果证明，这种策略是无效的。两位教授还用加德纳兄弟挖掘过的那段时间的数据对"傻瓜四股"策略进行了重新检验。这一次，他们做了一个巧妙的调整。他们没有选择 1 月第一个交易日的投资组合，而是在 7 月第一个交易日实施了这种策略。如果这种策略真的有价值的话，它不应该对起始月份具有敏感性。当然，两位教授发现了这种敏感性。

1997 年，就在提出"傻瓜四股"策略一年以后，加德纳兄弟对这个系统进行了改进，将其更名为 UV4。他们的解释证实了他们的数据挖掘："为什么要调整？历史证明，UV4 的表现优于之前的'傻瓜四股'。"通过数据挖掘得到的策略无法在编造理论所使用的年份之外取

得同样优秀的表现，这一点儿也不令人吃惊。当“彩衣傻瓜”在 2000 年停止推荐“傻瓜四股”和 UV4 时，加德纳兄弟承认了这一点。

“傻瓜四股”策略和它的名字一样愚蠢。

反向头肩

超级碗指标是预测股票市场上涨或下跌的简单规则。有一个行业专门通过测量投资者的情绪预测股票价格，这个行业叫做技术分析。基本分析用利润、利率以及其他经济因素衡量股票的价值。技术分析则认为经济因素是众所周知的事情，已经被考虑到了市场价格之中。观察基本因素是没有意义的。更加有用的做法是测量投资者的情绪，也就是凯恩斯所说的“动物精神”。

技术分析的核心是确定股价模式，以预测未来的价格。这些模式被贴上了标签，比如通道、支撑线、阻力线、双顶、双底、头肩、杯子和把手，这使分析看上去更具合理性。不过，一项又一项的研究发现，虽然技术分析拥有这些吸引人的标签，但它几乎没有任何价值——除了雇用技术分析师以及为股票经纪人带来佣金。

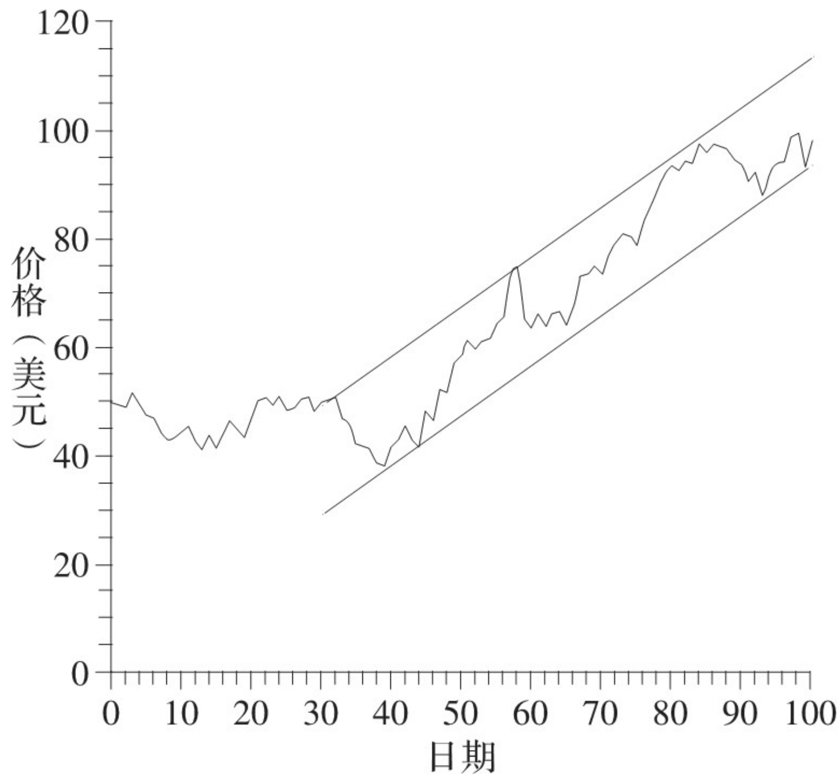


图 16.5　上升通道

一位经济学家曾向一位技术分析师（让我们叫他爱德吧）寄了一些股票价格图表，包括图 16.5，并请爱德帮他确定其中是否存在一些看上去适合投资的股票。经济学家没有指出公司的名称，因为最纯粹的技术分析师希望专注于价格模式，不想被自己对于一家公司所持有的信息或感觉左右。一些技师回避一切新闻，在没有窗户的房间里工作，因为他们不想被消息的好坏或天气的阴晴所影响。

分析师在图 16.5 上画了两条平行线，看到了明显的模式。大约从第 30 天开始，这只股票的交易进入了一条狭窄的倾斜上升通道。在第 100 天，价格接近通道的下边界，显然即将开始飙升。

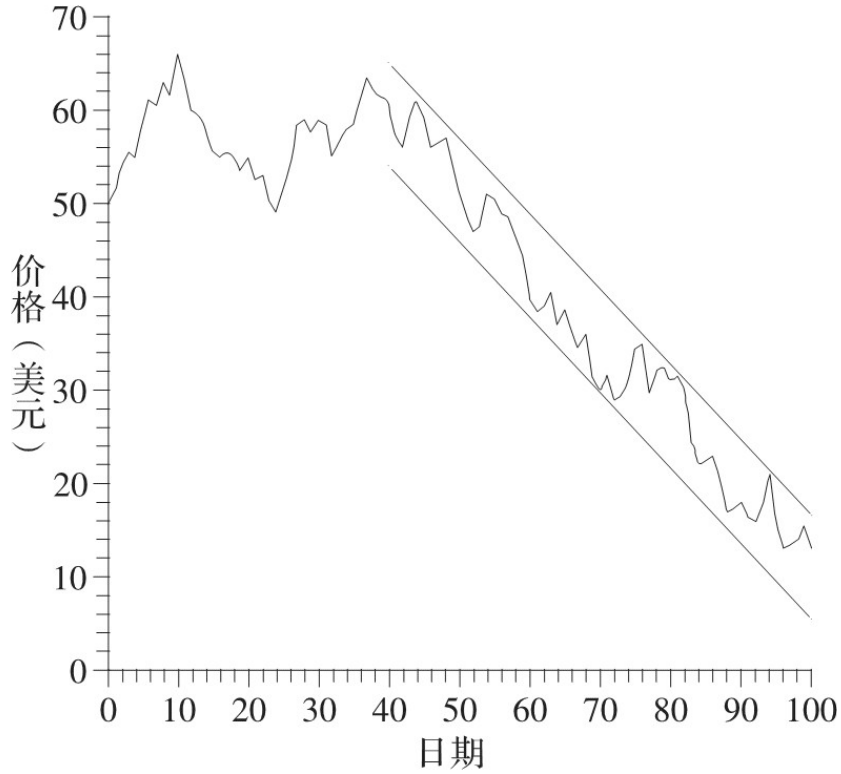


图 16.6　死亡螺旋

图 16.6 也具有明显的模式。这次的通道向下倾斜，股票显然进入了死亡螺旋。只有傻瓜才会购买具有这种历史价格的股票。

图 16.7 的股票拥有一条支撑线，这条线是在第 24 天形成的，随后得到了两次证实。每当价格下降到 28 美元时，它都会得到支撑，开始向上反弹。更重要的是，这张图显示了头肩模式：价格先是从 28 美元的支撑线上升，然后回到 28 美元，然后出现了幅度更大的上升，然后回到 28 美元，然后温和上升。技术分析师相信，通过头肩模式确立和证实的支撑线极为牢固。当价格在第 99 天击穿 28 美元支撑线时，这显然意味着出现了严重的问题。只有大量坏消息才能导致股价突破如此牢固的支撑线。一旦 28 美元的障碍被攻破，价格就只能向下走了。

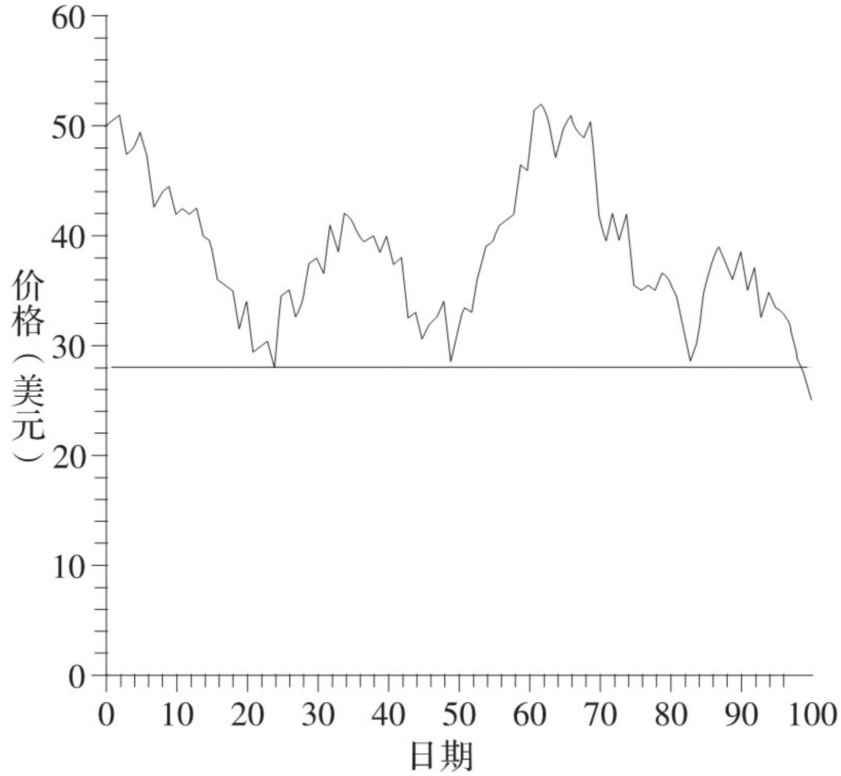


图 16.7　击穿头肩支撑线

图 16.8 显示了相反的模式。这张图确立和证实了 65 美元的阻力线。每当价格接近 65 美元时，它就会反弹。这种情况出现得越多，价格超越 65 美元的心理阻碍就越大。更重要的是，在技术分析师看来，这张图显示了反向头肩，因为第二次始于 65 美元的反弹比第一次和第三次反弹要强烈得多。当价格在第 98 天超越阻力线、消除价格进一步上升的心理障碍时，这显然是一个买入信号。

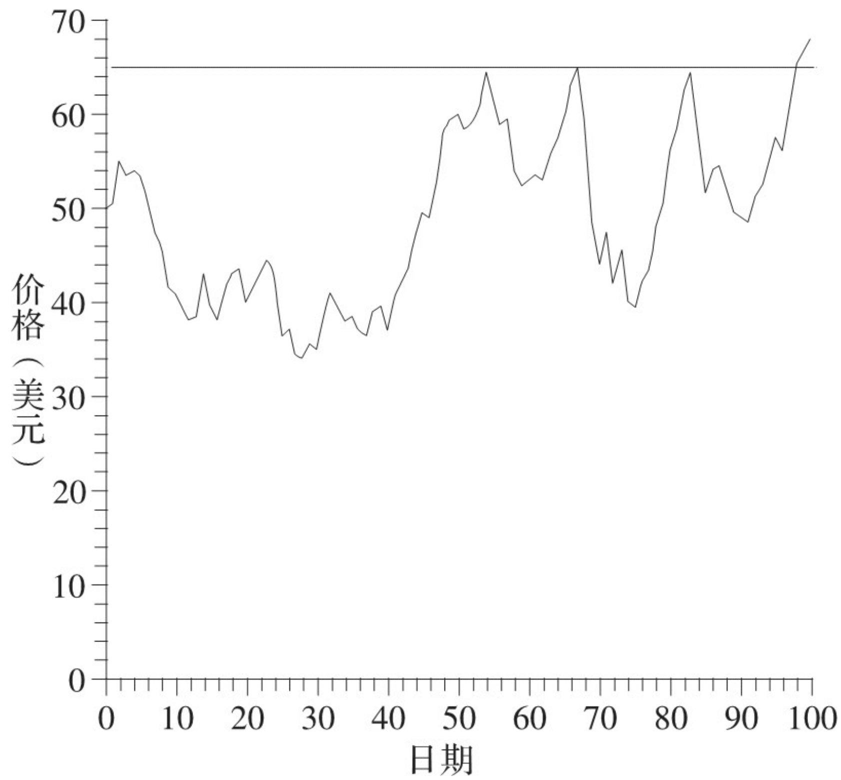


图 16.8　突破阻力线

分析师看到这些模式时非常激动，因此忽略了一个奇怪的巧合：所有四张图的价格起始点都是每股 50 美元。这并不是巧合。

这些图像并不是真实的股票。向爱德寄出这些图像的恶作剧教授（我承认，那就是我）根据学生抛掷的硬币编造了这些虚构的数据。在每张图上，“价格”始于 50 美元，每天的价格变化取决于 25 次抛硬币的结果。如果硬币正面朝上，价格将上升 50 美分；如果硬币背面朝上，价格将下降 50 美分。例如，如果抛出 14 个正面和 11 个背面，当天的股价将上涨 1.50 美元。在生成几十张图以后，我把其中的十张图寄给分析师，相信他能够找到具有吸引力的模式。当然，他做到了。

当我把真相告诉他时，爱德非常失望，因为他分析的不是真实的股票，没有机会通过买入和卖出赚取利润。不过，他从这个恶作剧中得到的教训与我的预想完全不同：爱德认为技术分析可以用于预测抛硬币的结果！

这个例子真正的教训是，即使是专业投资者也很难理解“数据挖掘将不可避免地使人发现完全出于巧合的统计模式”这个道理。缺乏理论的数据虽然诱人，但却具有误导性。

如何（不）中彩票

在纽约州乐透游戏中，玩家从 1 到 59 这 59 个数字中选择 6 个数字，如果它们与电视直播中选出的 6 个数字相同（不计顺序），玩家就能赢得大奖。获胜的可能性是四千五百万分之一，但一张彩票只要 50 美分。在 25 年时间里，大奖的奖金相当于彩票总销售额的 30%。另外 15% 的彩票收入被分配给小一些的奖项。在美国所有的六合彩游戏中，纽约乐透的支出是最低的，获奖概率也是最低的。例如，附近的康涅狄格州是 44 选 6，新泽西州是 49 选 6。不过，在美国六合彩游戏中，纽约乐透是最受欢迎的游戏。

各州的彩票实际上是一种智商税，但乐透游戏的流行说明许多人渴望获得彻底改变人生的机会——不管这种机会多么遥远。2012 年，覆盖美国大部分地区的“百万大博彩”提供了创纪录的 6.56 亿美元头奖。数千人排队几个小时或者开车 160 公里，以加入购票大军，他们一共购买了 15 亿美元的彩票。这场彩票热潮的唯一陷阱是，获得创纪录头奖的可能性极其微弱，只有一亿七千六百万分之一。

梦想与现实之间的巨大差距为一些具有创业精神的人提供了销售骗人伎俩的机会，他们声称这些方法可以提高中彩票的可能性。一家公司以 19.95 美元的价格销售“百万富翁生成器”，他们的广告语是：“研究表明，大多数乐透中奖者并没有使用特殊的策略来选择数字。相反，他们运用了随机选择的力量。”这个价值 19.95 美元的产品是一个由电池驱动的球体，里面装满了带有数字的小球，它可以“将小

球的顺序完全打乱”，并且可以选出“一组完全随机的数字”。我说过，彩票是一种智商税，没错吧?

另一些人提出了一种完全不同的方法。他们声称，通过收集和分析中奖数字，我们可以像投注赛马一样投注彩票。你很难想象出更加浪费时间的事情。也许，计算电话号码的平均值比它更加无聊?

盖尔·霍华德（Gail Howard）的网站声称，她是“30 年来美国最可靠的彩票权威”。根据该网站的说法，她还上过数百次广播和电视节目，包括《今日秀》和《早安美国》；在一年多的时间里，她的 30 分钟专题广告片《彩票克星》每天都在有线电视上播出。（奇怪的是，我认为这种吹嘘是一种负面宣传，而不是正面宣传。这种观点同样适用于她的警告：“你最好趁着这些具有争议性的信息仍然被允许出售的时候抓紧时间。”）霍华德的许多书籍以下面的陈述作为开头，这段陈述唯一值得注意的地方就是它完全否定了书中讲述的其他内容：“你不需要担心州级彩票游戏受到控制或操纵……中奖号码是通过完全随机的过程选出来的。”霍华德并没有因此而气馁，她提供了一些建议，用于“极大地提高你的中奖机会”。

下面是霍华德和其他专家提供的赢得六合彩乐透游戏的前六条秘诀：

> 1. 选择热门数字。纽约彩票网站显示了过去几年的中奖数字。将它们输入到电子表格里，找出最经常出现的数字。
>
> 2. 不要忽视迟迟不出现的数字。最近没有出现的数字很可能会突然出现。
>
> 3. 不要在连续数字上下注，比如 11、12、13、14、15、16。连续数字几乎没有中过奖。
>
> 4. 不要用计算机做出迅速的选择。计算机知道什么呢?
>
> 5. 男士赢得了三分之二的头奖。如果你是女士，请让你的丈夫、男友、兄弟或父亲为你购买彩票。

6. 中奖者最常见的名字是玛丽（Mary）和约瑟夫（Joseph），包括玛丽亚（Maria）和乔（Joe）等变体。如果你拥有与此不同的名字，请让你的朋友玛丽或约瑟夫为你购买彩票。

下面是一个持有怀疑态度的人对这些忠告的回应：

1. 每个数字被选中的可能性是相等的，除非游戏受到操纵（没有人正式做出这种论断）。选择中奖数字的设备既不会记得过去抽到的数字，也不会关心未来抽到的数字。

2. 这种说法不具有指导意义。你刚才让我在热门数字上下注，现在又让我在冷门数字上下注。

3. 没错，连续数字不太可能中奖。同理，任何六个数字都不太可能中奖。六个数字的每一种组合具有同样微弱的中奖机会。

4. 没错，计算机无法预测中奖数字。同理，你也无法预测中奖数字。这就是它被称为“随机”数字的原因。

5. 也许男士购买了三分之二的彩票。或者，也许女性中奖者让她们的丈夫、男友、兄弟或父亲替她们领奖。

6. 也许，仅仅是也许，彩民中最常见的名字是玛丽和约瑟夫，包括玛丽亚和乔等变体。

正如仔细的研究可以在抛硬币和掷骰子的结果中发现毫无价值的模式，耐心的观察也可以在彩票数字中分辨出毫无价值的模式。对于彩票秘籍最有说服力的怀疑理由是，如果有人拥有一种切实有效的秘籍，他就会通过购买彩票致富，而不是兜售图书和电池驱动型设备。

没有证据表明盖尔·霍华德中过彩票。实际上，霍华德是本·巴克斯顿（Ben Buxton）创造出来的，后者以盖尔·霍华德、美国占星术协会、鲁迪哥哥（Brother Rudy）、都德夫人（Madame Daudet）、罗斯柴尔德受托人和卢尔德水十字架等几十个法人的名义经营着各种骗人的邮购生意，兜售彩票建议、占星预测术和好运片。

美国纽瓦克市检察官请求联邦地方法院禁止巴克斯顿“继续通过自称能够为天真的消费者带来财富、好运和成功的虚假骗人的天宫图、彩票秘籍、宗教物件和其他产品欺骗数十万消费者”。美国邮政部门也参与进来，和检察官共同争取到了一条法院命令，永久性地终止巴克斯顿的各种生意，使邮政部门有权截取寄给巴克斯顿的数万张支票，并且要求巴克斯顿向之前的消费者支付 30 万美元的赔偿。

不会破裂的泡沫

能够让我们与钱财告别的不仅仅是骗子。我们有时还会亲自参与到自我伤害的集体骗局之中。没有互联网的生活——无法用指尖访问电子邮件、亚马逊和维基百科的日子——是难以想象的。在停电或度假时，离开互联网的痛苦可能会令人难以忍受。手机进一步加重了我们的网瘾。我们真的需要每周 7 天、每天 24 小时在线并随时接听手机吗？我们必须立即回复每一封电子邮件以及每一条信息和推文吗？我们真的需要知道我们所有的朋友午餐吃的是什么吗?

早在 20 世纪 90 年代，当计算机和手机刚刚开始进入我们的生活时，互联网的发展催生出了数百家基于互联网的公司，通常被称为“.com”公司。一些网络公司拥有良好的理念，发展成了强大而成功的公司。不过，许多公司并没有做到这一点。很多时候，人们只想开一家名字里带有“.com”的公司，将其卖给别人，然后带着大量钞票离

开。产生不错的想法、创办一家公司、使其成为成功的企业、将其传给儿子和孙子的经济思想实在是太陈旧了。

一家网络公司证明自己的方式不是赚取利润，而是花钱，最好是花别人的钱。（我没有开玩笑！）一种理由是，公司应该迅速壮大，成为行业领头羊。（有一种流行的说法，叫做"不壮大就会消失。"）这种观点认为，当人们相信你的网站是购买某样东西、销售某样东西或者学习某样东西的地方时，你就实现了垄断，可以消除竞争，收获利润。

这并不完全是一种愚蠢的想法。有时，它甚至是有效的。（想一想亚马逊和 eBay。）问题是，就算垄断某件事情是可能的，别忘了，美国有数千家网络公司，但是不可能出现数千个垄断者。在试图迅速壮大的数千家公司中，只有极少数公司能够实现垄断。

大多数网络公司没有利润。因此，心怀梦想的投资者为所谓的"新经济"想出了新的衡量方法，以支持不断上涨的股票价格。他们说，我们不应该沉迷于像利润这样古老的事物；相反，我们应该考察一家公司的销售额、支出和网站访问量。作为回应，各家公司通过具有创意性的方式向投资者提供他们想要的东西。投资者想要更高的销售额?我把某样东西卖给你们公司，你再把它卖给我。我们没有赚到任何利润，但我们的销售额都在提升。投资者想要更多的支出？再去订购一千把艾龙椅子。投资者想要更高的网站访问量？向访问你们网站的人发放小礼品。购买超级碗广告，宣传你们的网站。记住，投资者想要的是网站访问量，而不是利润。24 家网络公司在 2000 年 1 月的超级碗比赛期间投放了广告，每 30 秒广告的成本是 220 万美元，这还不包括制作广告的费用。公司不需要利润。它们需要的是流量。

从 1995 年到 2000 年，股票价格增长了两倍，年增长率为 25%。网络公司的股票涨幅更大。以科技股著称的纳斯达克指数在 5 年时间里增长了四倍，年增长率为 40%。如果一个人幸运地在 1995 年 1 月购买了 1 万美元的美国在线股票，或者在 1996 年 4 月雅虎上市时购买了 1 万美元雅虎股票，那么到 2000 年 1 月，他将拥有将近 100 万美

元。股市投资者和网络公司创办者都在变富，他们相信这个过程永远也不会结束。当然，事情终有结束的那一天。

2000 年 3 月 11 日，我参加了一场座谈会，会议的议题是蓬勃发展的股票市场以及受到广泛宣传的“36K”预测：道琼斯工业平均指数将增长两倍多，从不到 12000 点增长到 36000 点。第一位发言者谈论了摩尔定律（集成电路中的晶体管密度每两年翻番）。我听得很认真。我必须承认，科学技术的发展速度令人吃惊。不过，发言者对于“股票价格过高、过低还是刚刚好”的问题只字未提。

下一位发言者谈论了网络公司的专家们多么聪明。当你购买网络公司的股票时，你把钱交给了非常聪明的人，他们很快就会找到某种非常赚钱的资金处理方式。我仍然听得很认真。我必须承认，许多网络公司是由讨人喜爱的聪明人创办的。你猜怎么着，我的一个儿子和其他四个刚毕业的大学生创办了一家公司。他们五个人在新罕布什尔租了一幢带有五间卧室的房子（税率不高，但却靠近“苛税诸塞州”），睡在楼上，每天通过楼梯前往办公场所。他们在楼下做什么工作呢？他们并没有商业计划。他们的关键词是“敏捷”。他们聪明、灵活、具有创造性。当一个赚钱的机会出现时，这五个小伙子就会敏捷地意识到这一点，并用十只手抓住这个机会。我承认，他们是很棒的孩子；而且，这个世界上显然还有几百个很棒的孩子正在寻找通过互联网盈利的途径。不过，发言者对于“股票价格过高、过低还是刚刚好”的问题仍然只字未提。

下一位发言者谈论了美国负责货币政策的政府机构联邦储备委员会（美联储）主席艾伦·格林斯潘（Alan Greenspan）多么优秀。美联储负责决定何时增加货币供应以刺激经济，何时限制货币供应以减少通胀压力。一位犬儒主义者（我）曾经写道，每当美联储觉得失业最符合我们的利益时，它就会提高利率，导致经济衰退。拥有一位知道自己正在做什么的联储主席是非常重要的。我听得很认真。我必须承认，艾伦·格林斯潘是一位令人印象深刻的联储主席。不过，发言者对于“股票价格过高、过低还是刚刚好”的问题仍然只字未提。

我是最后一名发言者，也是这场欢乐的聚会上唯一提出消极观点的人。我从各种角度对股票价格进行了考察，认为“道指有可能在某个时候达到 36000 点”的想法是靠不住的，而且当前的股票价格高得离谱。我的结束语是，“这是一个泡沫，它很快就会发生严重的破裂。”

可怕的预言变成了现实。这场会议的日期是 2000 年 3 月 11 日星期六。纳斯达克指数在接下来的星期一开始下跌，在三天时间里从 2000 年 3 月 10 日的峰值下跌了 75%。美国在线下跌了 85%，雅虎下跌了 95%。最有趣的问题不是我的评论在时间上的巧合，而是为什么我相信这是一个泡沫。

在网络公司的泡沫中，大多数人没有使用令人信服的理论去衡量股票价格过高、过低还是刚刚好。相反，当他们看到股票价格上涨时，他们编造出了一些理由，以解释正在发生的事情。他们谈论摩尔定律、聪明的孩子以及艾伦·格林斯潘。这是缺乏理论的数据。

下面是一个言之有理的理论。我们可以将股票想象成每年提供红利（比如 2 美元）的赚钱机器。这台优秀机器的经济价值是你为了每年获得 2 美元红利而愿意支付的价格。

投资者购买赚钱机器是为了每年获得红利。相比之下，投机者购买股票是为了通过卖给其他人赚取利润。对投机者而言，股票的价值是其他人愿意支付的价格，他们要做的是猜测其他人明天愿意为你今天购买的东西支付怎样的价格。这种猜谜游戏就是所谓的“博傻理论”：以虚高的价格购买某样东西，希望能够以更高的价格将其卖给更大的傻瓜。

在投机泡沫中，赚钱机器的价格远远超出了其经济价值，因为人们购买这台机器不是为了每年的 2 美元，而是为了以更高的价格将机器卖给其他人。他们之所以认为未来的价格会上升，仅仅是因为过去的价格出现了上升。（还用我说吗？这仍然是缺乏理论的数据。）当投机者不再认为这台机器的价格会不断上升时，泡沫就会破裂。他们开始销售机器，价格开始不断下跌，因为如果投机者认为他们无法以

更高的价格将机器卖出去，他们就不会支付虚高的价格。当他们失去信心时，这场盛宴也就结束了。

如果股票的经济价值是 100 美元，市场价格是 500 美元，这就是一个泡沫，因为股票价格远远超出了其经济价值。你觉得某样东西的价格永远不可能如此偏离经济价值？在 17 世纪，荷兰郁金香球茎的售价是几万美元（以今天的美元计算）。最近，毫无价值的填充娃娃可以卖到几百美元。

豆豆娃是由泰·沃纳（Ty Warner）制造的填充动物玩具，带有心形吊牌，里面装有塑料小球（“豆豆”）。在 1995 年左右，就在网络公司泡沫开始膨胀的同一时间，豆豆娃开始被视为“收藏品”，因为买家认为他们可以将这些笨熊卖给无穷无尽的傻瓜，以便从不断上涨的价格中获利。一些陷入幻想中的个体开始囤积豆豆娃，认为它们可以提供退休金或者孩子的大学学费。

豆豆娃到底有多大的经济价值呢？它无法提供红利。它无法提供任何报酬！你甚至无法用它来玩。要想保持豆豆娃的收藏价值，你必须将它存储在密封的容器里，放置在凉爽、阴暗、无烟的环境中。不过，满怀希望的人和贪婪的人愿意为这种最初在玩具店里只卖几美元的豆豆娃支付几百美元。他们看到了豆豆娃的价格在过去的大幅增长，认为未来也将如此。他们没有理由相信这一点，但他们希望自己相信。这仍然是缺乏理论的数据。

1999 年，纪念威尔士王妃戴安娜（Diana）的王妃豆豆娃开始发售。到 2000 年，这款豆豆娃已经卖到了 500 美元。接着，泡沫破裂了。2008 年，我在亚马逊买了一个王妃小熊，其运输费用已经超过了小熊本身的价格。

南海泡沫

网络公司泡沫和豆豆娃泡沫并非新鲜事物。1720 年，英国政府为南海公司提供了西班牙美洲殖民地的独家贸易权。公司的董事从未去过美洲，他们也没有任何具体的贸易计划，但在公司具有创造性的记账方式的刺激下，人们纷纷为这个海外项目投资。南海公司的股价从 1 月 28 日的 120 英镑开始飙升，5 月 19 日达到了 400 英镑，6 月 4 日达到了 800 英镑，6 月 22 日达到了 1000 英镑。一些人成了暴发户，数千人希望挤进这个行列之中。据说，你可以在走进卡洛韦咖啡馆时购买股票，并在离开时将其卖掉，以赚取利润。人们看到了模式——以某个价格买入，以更高的价格卖出——但他们并没有想过这种模式是否合理。

其他一些头脑胜过良知的企业家开始以更加宏大的计划发售股票，引来了大量不想掉队的疯狂投资者。几乎任何计划都会受到追捧。一家公司承诺制造永动轮。另一家公司的成立目的是“开展一项具有巨大优势的事业，但是没有人知道这个事业是什么”。（就像某些网络公司一样！）这家神秘公司的股价被定为 100 英镑，并且承诺每年回报 100 英镑。公司的发起人在 5 个小时之内卖掉了所有股票，然后立即离开英国，再也没有回来。还有一家公司的股票是为了“尼特温德”，即不销售任何东西。当南海泡沫破裂时，财富和梦想一齐消失了。

和其他所有投机泡沫一样，许多人相信博傻理论。一些人怀疑价格不合理，但是市场上的大多数人都相信价格还会继续上升，至少能够持续到他们将其卖给下一个排队的傻瓜以后。1720 年春天，艾萨克·牛顿爵士（Sir Isaac Newton）说：“我能计算天体的运动，但我无法计算人类的疯狂”。他卖掉了南海股票，赚了 7000 英镑。当年晚些时候，就在泡沫破裂之前，他再次购买了股票，并且损失了 2 万英镑。一位银行家在南海股票第三次发售时投资 500 英镑，他解释说：“当世界上的其他人陷入疯狂时，我们必须以某种程度模仿他们。”英国议会成员詹姆斯·米尔纳（James Milner）在南海泡沫中破产，他悲叹道：“我说，的确，我们一定很快就会遇到灭顶之灾，但是……它比我的预测早来了两个月。”

伯克希尔泡沫

一些投资者认为，他们可以查看某种事物的价格相对于消费者物价指数（CPI）的增长速度，以确定是否存在泡沫。根据这种没有任何理论支撑的数据，他们说，如果资产价格上升得很快，说明一定存在泡沫。事实未必如此。缺乏理论的数据总是不可靠的。

大多数优质股票每股的价格在 20 美元到 200 美元之间。2005 年，富国银行的股价是 30 美元左右，可口可乐是 40 美元，强生是 60 美元。令人吃惊的是，伯克希尔·哈撒韦的股价是 9 万美元。众所周知，伯克希尔是由传奇投资者沃伦·巴菲特经营的，但是一股股票的价格怎么能超过一辆保时捷 911 呢?

图 16.9 对 1995 年到 2005 年伯克希尔股价的增长百分率与消费者物价指数的增长百分率进行了比较。消费者物价指数增长了 27%，伯克希尔则增长了令人震惊的 269%。伯克希尔是一个即将破裂的泡沫吗?

此后，伯克希尔的股价继续增长，在 2007 达到 141600 美元峰值，2009 年和其他股票一同下跌，随后向上蹿升，2013 年达到了 17 万美元以上。根据图 16.10，2005 年以后，伯克希尔与消费者物价指数之间的实际差距出现了增长。

一股股票的价值怎么能超过一辆保时捷呢？股票价格的增速怎么能超过消费品价格（食品、服装和住房的价格）呢?

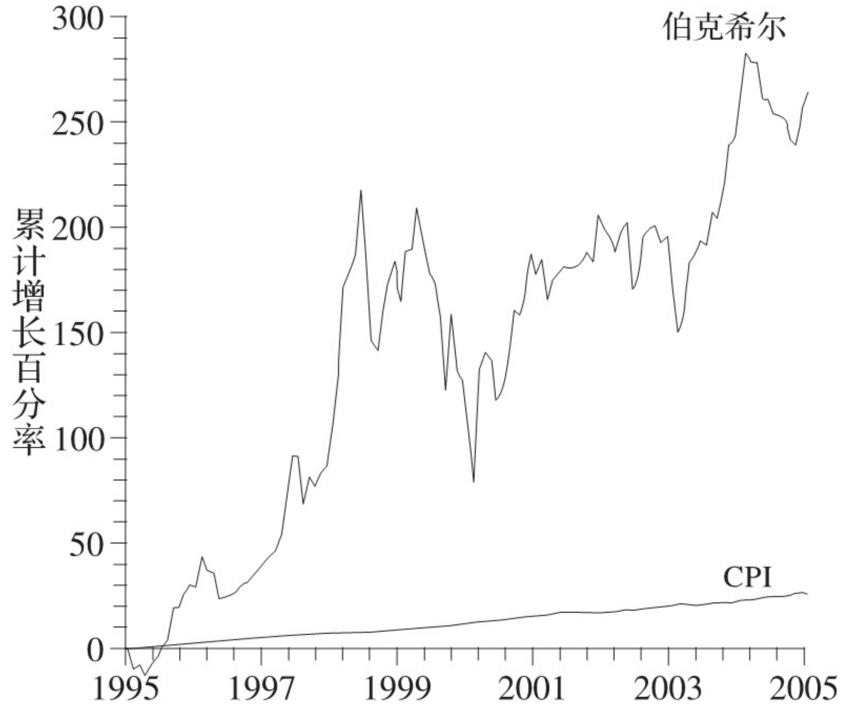


图 16.9　消费品物价指数和伯克希尔·哈撒韦股价的增长百分率



图 16.10　消费者物价指数和伯克希尔·哈撒韦在更大的时间尺度上的增长百分率

公司是由股东所有的，因此股票的价值取决于公司的价值。进一步说，一张股票的价值取决于公司的整体价值以及股票的发售数量。如果一家公司的总体价值是 1 亿美元，并且发售了 1 百万股股票，那么每股的价值是 100 美元；如果公司发售了 1 亿股股票，那么每股的价值只有 1 美元。

许多公司通过股份分割保持股票的“可购买性”。假设 Things.com 的价值是 1 亿美元，并且发售了 1 百万股股票，那么每股的价值是 100 美元。经过多年盈利，公司的价值增长到了 2 亿美元，每股的价值是 200 美元。Things.com 可以进行二比一股份分割，使股票数量翻番，达到两百万股，同时每股价值折半，变成 100 美元。股份分割是一种没有做出任何改变的举动，每个股东的股票数量增长了一倍，但是每股的价值减少了一半。

伯克希尔的股票之所以如此昂贵，是因为它从不分割，同时伯克希尔在巴菲特的指导下出现了巨大的升值。巴菲特总是认为股份分割没有意义，他有一个著名的生日贺卡签名：“祝愿你活到伯克希尔分割股份的那一天。”伯克希尔发售的股票不到两百万股。IBM 拥有超过 10 亿股股票，埃克森美孚拥有 60 亿股股票。这就是一股伯克希尔股票的价值超过一辆保时捷的原因。

关于第二个问题——股票价格的增速怎么会超过消费品价格——答案也很简单。一家公司的价值与一块肥皂的价格存在根本区别——即使这是一家销售肥皂的公司。公司的价值取决于利润，而利润不仅取决于肥皂的价格，也取决于肥皂的销量。

伯克希尔·哈撒韦拥有几十家公司，包括本杰明摩尔、北伯灵顿、冰雪皇后、鲜果布衣、盖可以及喜诗糖果。伯克希尔还在其他许多公司拥有很大的持股比例，包括美国运通、可口可乐、卡夫、宝洁、华盛顿邮报以及富国银行。伯克希尔的价值投资带来了大量收入，这些收入又被它投入到更多的公司之中。在将近 50 年的时间里，伯克希尔的总价值一直保持着每年超过 20% 的增速。

伯克希尔能够拥有超过 10 万美元的股价和超过消费者物价指数的增速，而且不构成泡沫，是因为这家不同寻常的公司拥有不同寻常的利润。

真实股价

扣除通胀因素后的数据被称为真实数据。如果我们的收入提高 10%，消费品价格也提高 10%，那么我们的真实收入没有发生变化，因为我们能够买到的东西和之前相同。

1987 年，两位商学院教授用标普 500 股价指数和消费者物价指数（CPI）计算了 1857 年到 1985 年的真实股价。他们发现，1985 年的真实股价高于 1857 年；股票价格的增长超过了消费品价格。

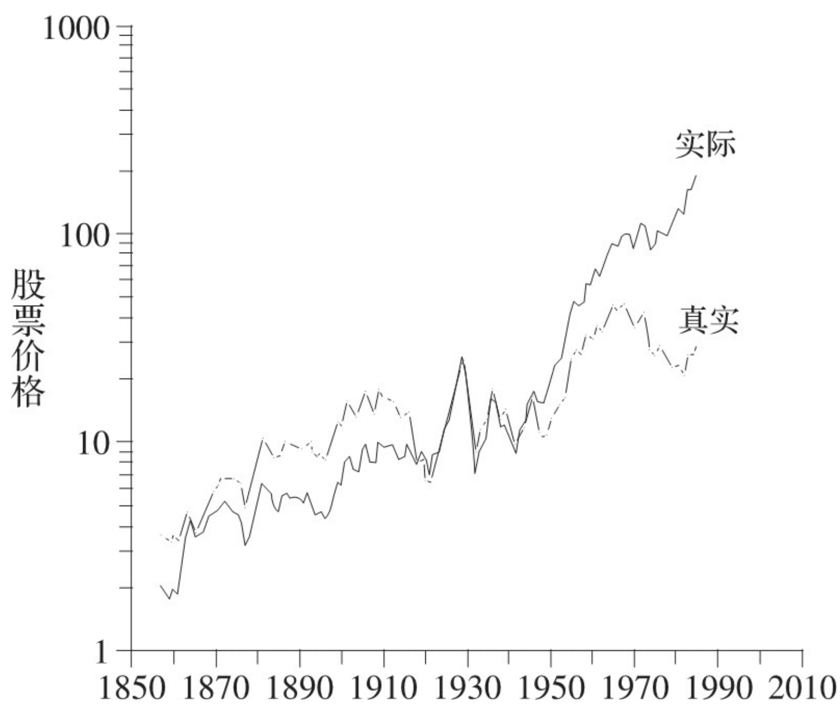


图 16.11　另一个泡沫

这不是什么新奇的观点。两位教授希望找到某种原创内容，某种可以发表的内容。因此，他们将实际股价与真实股价进行了比较，如图 16.11 所示，尽管这种做法看上去似乎没有任何意义。

两位教授将 1864 年的消费者物价指数置为 1，因此这一年的真实股价与实际股价相等。1864 年以后，实际股价跌到了真实股价之下。接着，两条线在 1918 年重合。从 1918 年到 1946 年，实际价格与真实价格共同运动。接着，二者再次分开，这意味着股价应当下跌，以便使两条曲线再次重合。第二次世界大战以后股价的飙升显然是一个无法持续的泡沫。

就像两位教授希望的那样，这种比较的原创性使它得到了发表。问题是，实际股价与真实股价的比较属于缺乏理论的数据，它没有任何意义。在比较两种事物之前，我们应当思考二者之间是否可能存在联系。实际股价与真实股价的比较有些类似于"伯克希尔的股票存在泡沫，因为其价格增速超过了消费者物价指数"这一错误观点。不过，基于图 16.11 的预测与上述观点存在很大的差异，而且更加异想天开。实际股价与真实股价之间的差异与股票价格没有任何关系！

别忘了，真实股价是实际股价扣除通胀因素后的产物。二者在 1864 年相等，因为 1864 年的消费者物价指数被设置成了 1。1865 年到 1918 年的真实股价高于实际股价，因为消费品的价格出现了下降。真实股价和实际股价在 1918 年和 1946 年之间相等，因为这段时间的消费品价格与 1864 年（近似）相同。1946 年以后，真实股价低于实际股价，因为消费品价格高于 1864 年的价格。这就是图 16.11 显示的全部内容。同 1864 年相比，消费品价格在 1865 年到 1918 年出现了下降，在 1946 年以后出现了上升。

"真实股价将与实际股价再次重合"的预测不是关于股市的预测。它是在预测消费者物价指数将回到 1864 年的水平。

图 16.12 说明，这种隐性预测是完全错误的。股市经历了令人欣喜的上涨和可怕的下跌，但实际股价和真实股价之间的差距仍然在逐

年扩大，因为消费品的价格每年都在提高。这种差距永远不会消失，因为消费品的价格永远不会再次回到 1864 年的水平。

这个例子的核心问题是，两位教授在比较实际股价和扣除通胀因素后的股价时并没有考虑到这种比较的意义。这种比较只能告诉我们消费品的价格是否高于 1864 年的水平——而我们已经知道了这个问题的答案。

木匠们常说一句话："两次测量，一次切割。"关于数据，这句话应该改成"两次思考，一次计算"。

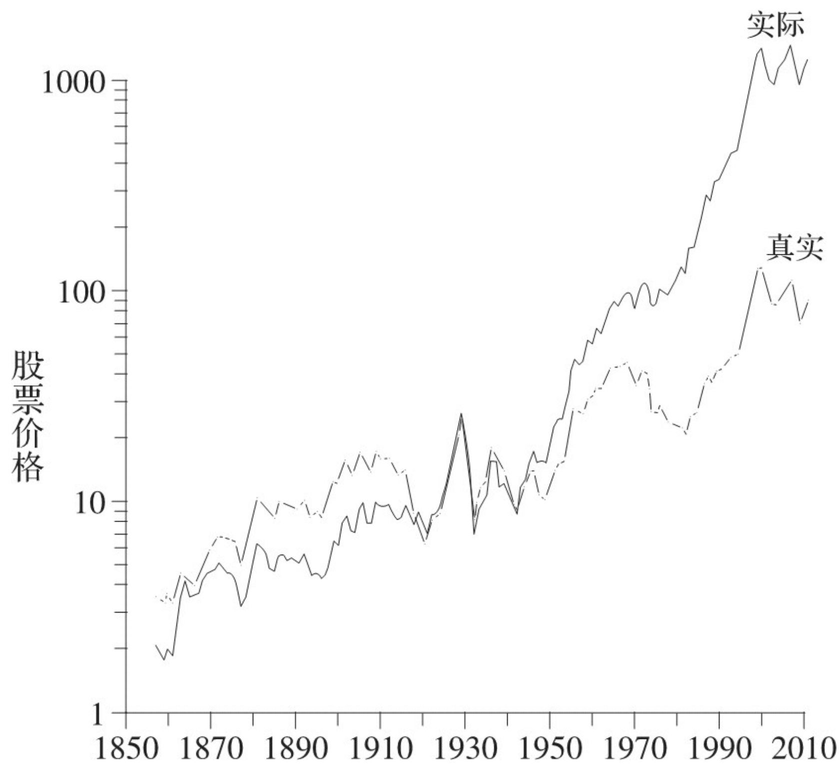


图 16.12　差距持续扩大，因为消费者物价指数一直在增长

×如何轻松识破一本正经地胡说八道

缺乏理论的数据可能导致错误推断。例如，通过将短短几年的数据外推到未来，人们可以编造出许多令人欣慰或恐

慌的预测。在感到欣慰或恐慌之前，考虑这种外推是否合理。仅仅通过观察过去预测未来的做法是否具有令人信服的理由？这是不是一种一相情愿的想法？或者根本没有任何思想？

数据中的模式也是同样的道理。别忘了，即使是随机抛掷硬币，也会产生奇特甚至令人震惊的模式，但它们是毫无意义的。当某人向你展示某种模式时，不管这个人的履历多么令人震撼，你都应该考虑这种模式仅仅出于巧合的可能性。问一问“为什么”，而不是“什么”。不管遇到怎样的模式，你都应该思考：为什么会出现这种模式？

如果两种事物之间没有合理的联系，那么对二者的统计性比较也是没有说服力的。为什么股票价格的增长应该与消费品价格同步？为什么扣除通胀因素后的股价应该等于未经调整的股价？二者之间可能仅仅存在一种虚假的相关性。问问你自己，开展这项研究的人在计算之前是否进行了思考。他们在考察数据之前是否拥有具体而清晰的理论？这种理论是否受到了未经污染的数据的检验？或者，数据是否受到了挖掘和搜刮？

第 17 章
超级投资者

20 世纪 80 年代，一家拥有“休姆联合公司”这一复杂名称的投资咨询公司推出了《超级投资者文件》。根据遍布全国的广告宣传，普通投资者可以利用这些复杂的策略收获不同寻常的利润。公司每月向订阅者邮寄小册子，每份小册子大约 50 页，以时髦的方式打印在厚纸上。一本小册子的价格是 25 美元，外加 2.50 美元运输和处理费用。

现在回想起来，一个显而易见的事实是，如果这些策略真的像公司宣传的那样赚钱，那么这家公司就会利用这些策略赚取更多的利润，而不是去销售什么小册子。不过，贪婪的投资者天真地认为，他们可以用 25 美元外加 2.50 美元运输和处理费用买到成为百万富翁的秘诀。

一种“超级投资者”策略基于金银比“GSR”，即一盎司①黄金与一盎司白银的价格比率。1985 年，黄金的平均价格是 317.26 美元，白银的平均价格是 5.88 美元，所以金银比为 317.26 美元/5.88 美元=54，这意味着一盎司黄金与 54 盎司白银的价格相等。

1986 年，休姆公司写道：

> 过去七八年，金银比的波动很大，1980 年降至 19 比 1，1982 年升至 52 比 1，1985 年升至 55 比 1。不过，你也可以

清晰地看到，它总是——总是——回到 34 比 1 和 38 比 1 之间。

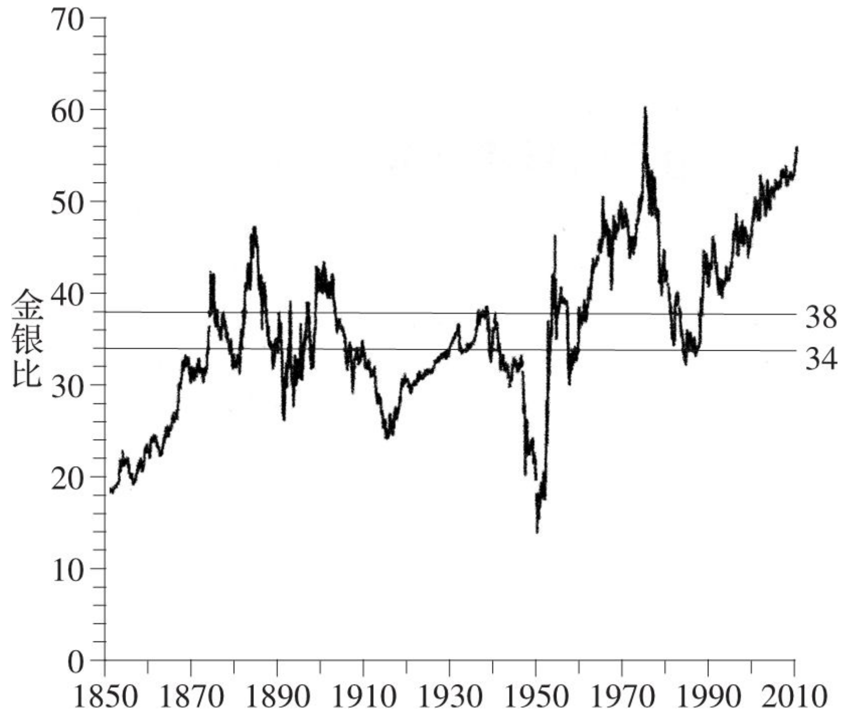


图 17.1　金银比，1970 ～ 1985 年

图 17.1 表明，1970 年到 1985 年的金银比的确在 34 到 38 这个区间周围波动。

这个“超级投资者”策略建议投资者在金银比高于 45 或低于 25 时投入资金，因为金银比总会回到 34 到 38 之间。例如，当金银比在 1984 年超过 45 时，金价相对于银价处于很高的位置。精明的投资者应该卖出黄金（因为它的价格很高），买入白银（因为它的价格很低）。

普通投资者没有可以用于销售的大量黄金，他们也不想在地下室囤积白银，但他们可以买入和卖出期货合约，即关于未来黄金和白银价格的赌注。卖出黄金期货是在押宝黄金价格下跌；买入白银期货是

在押宝白银价格上涨。同时卖出黄金期货和买入白银期货是在押宝金银比下跌。

1984 年 7 月 3 日，金银比为 45.04，金价为 369.75 美元，银价为 8.21 美元。假设你买入一份白银期货合约，卖出一份黄金期货合约，期待金银比回到 34 和 38 之间。

进一步说，假设你的预测是正确的。黄金涨到了 380 美元（涨幅 2.8%），白银涨到了 10 美元（涨幅 22%），导致金银比降至 38。一份黄金期货合约对应于 100 盎司黄金；一份白银期货合约对应于 5000 盎司白银。所以，你的利润是 7925 美元：

白银收益	5000×（10.00 美元-8.21 美元）	=	8950 美元
黄金损失	100×（369.75 美元-380.00 美元）	=	-1025 美元
净利润			7925 美元

期货市场实行每日结算制，当天亏损的投资者需要把资金转移给获得收益的投资者。经纪人要求客户每天开盘时账户里拥有超出当天潜在结算金额的资金（叫做“保证金”）。如果你没有钱，经纪人就会对你的期货合约进行平仓处理。

假设你的保证金是 2500 美元。在这种情况下，2500 美元投资将获得 7925 美元利润，投资回报率达到了令人垂涎的 317%。

没错，金银比的温和下降为你带来了 317% 的投资回报。这是因为你使用了巨大的杠杆。你投入了 2500 美元，但你用来收获利润的期货合约价值约为 40000 美元。杠杆不是无中生有的财富。杠杆是众所周知的双刃剑，既可以放大收益，也可以放大损失。事实上，1984 年的金银比并没有迅速回落到 38 以下。相反，损失来得异常迅速。1984 年 7 月 6 日，在开始这项交易三天以后，金价为 366 美元，银价为

7.60 美元，金银比高达 48.16。你的仓位损失了 2675 美元，超出了你的全部初始投资。你要么追加保证金，要么平仓。这就是杠杆的威力。你可以在三天时间里失去全部投资。

如果你投入更多保证金，希望金银比回落，你的损失将会月复一月、年复一年地增长。图 17.2 表明，金银比此后持续增长，在 1986 年达到峰值 75，然后在 1987 年回落到 50。接着，金银比再次增长，1991 年达到了 100。1998 年，金银比短暂降至 45 以下，随后再次增长。直到 2011 年，即建立金银仓位 37 年以后，金银比才会回到 38 以下。从 1970 年到 1985 年，金银比的平均值是 36；从 1986 年到 2012 年，金银比的平均值是 66。

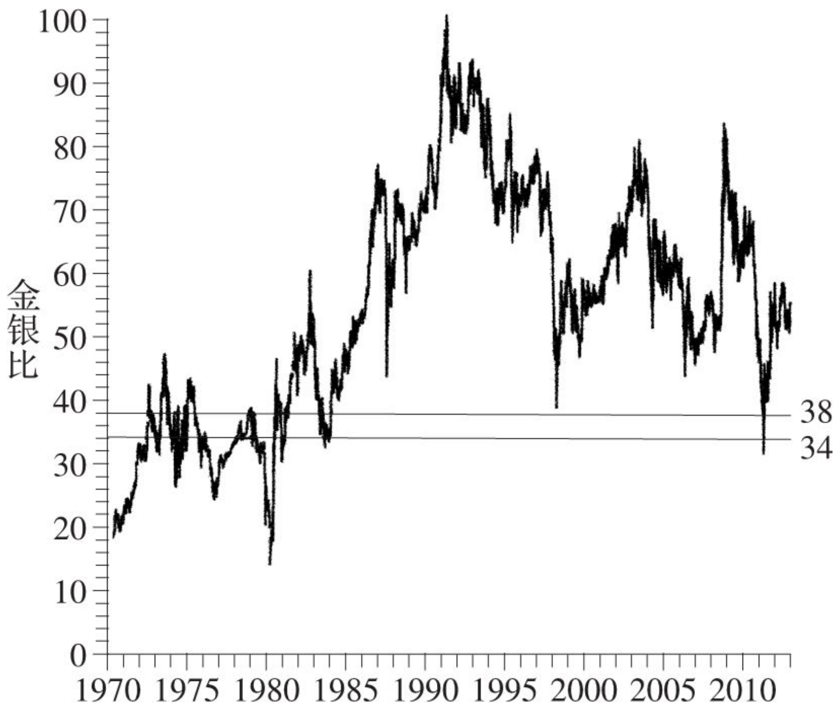


图 17.2　金银比，1970 ～ 2012 年

投资者很少能够将仓位维持 37 年，在每次期货合约更新时支付佣金，并在每个亏损日追加保证金。当金银比 1986 年达到 75 时，你的 2500 美元投资损失了 17000 美元。当金银比 1991 年达到 100 时，你

的损失是 22500 美元。如果你最初将更多资金（比如 25000 美元）投入到这项灾难性的交易中，你的损失就会变成原来的 10 倍。

宽客

计量金融分析师（“宽客”）用数学和数据来分析股票和其他投资品。他们不关心对总裁人格或产品潜力的主观评价。不要试图和他们谈论史蒂夫·乔布斯（Steve Jobs）、沃伦·巴菲特、智能手机或可口可乐。他们的口号是：“我只看数字！”

这种方法的先驱有哈里·马科维茨（Harry Markowitz）、罗伯特·默顿（Robert Merton）、费舍尔·布莱克（Fischer Black）、迈伦·斯科尔斯（Myron Scholes）等。马科维茨、默顿和斯科尔斯凭借自己的开创性理论获得了诺贝尔奖（布莱克已经去世），但他们的模型被人误用，这使许多投资者付出了巨大的代价。

20 世纪 50 年代，马科维茨提出了量化风险的均值方差分析。这种方法得到了机构投资者和财务咨询师的广泛使用。均值方差分析的一个主要弱点是对历史数据的依赖。投资者很容易假定过去相对安全的股票未来也是相对安全的。这是一个危险的假设！

20 世纪 70 年代，默顿、布莱克和斯科尔斯确定了看涨期权的理论“正确”价值（看涨期权的所有者有权在指定日期以指定价格购买股票）。这个模型目前被称为“布莱克-斯科尔斯”模型。它的一个主要弱点是，模型中的许多基本假设是错误的，比如股票价格的变化类似于抛硬币，投资者对股票和期权的交易没有成本，不会造成巨大而突然的价格变动。

这些人的开创性工作很好地说明了计量金融分析的两个主要陷阱：天真地相信历史模式是对未来的可靠指引，并且依赖于在数学上很方便却不切实际而且非常危险的理论假设。

最近出现的统计套利是计量金融的一种极端表现形式。和计量金融一样，它在数学和统计上很精致，但也很危险，因为它所依据的经验假设和理论假设并不可靠。

收敛交易

金银比是统计套利的一个早期案例。统计套利指的是所谓的“宽客”寻找某种统计模式，并且假设对这种模式的偏离是暂时的失常，可以加以利用。如果黄金与白银的价格比率历史上位于 34 到 38 之间，他们就会认为，如果金银比离开这个范围，那么它很快就会回归正常水平。这就是所谓的“收敛交易”。

投资者赌的不是价格上涨或下跌，而是价格收敛到某种历史关系。虽然价格对历史关系的偏差可能很小，但杠杆押注可以将很小的收敛转变成很大的利润。

早期的收敛交易基于简单的模式，比如金价与银价的比率。现代计算机可以对巨大的数据库进行搜索，寻找更加细微、更加复杂的模式。不过，根本的问题并没有改变。缺乏理论的数据是靠不住的。如果人们发现的模式没有深层次原因，那么我们没有理由认为对这种模式的偏离能够得到自我校正。

没有任何合乎逻辑的理由能够解释一盎司黄金的价格与 34 到 38 盎司白银的价格相等的原因。黄金和白银并不是一打和半打鸡蛋；如果后者的价格出现偏差，消费者就会购买更便宜的鸡蛋。黄金和白银也不是玉米和大豆；如果玉米价格相对于大豆价格出现上涨，农民就会种植更多玉米。

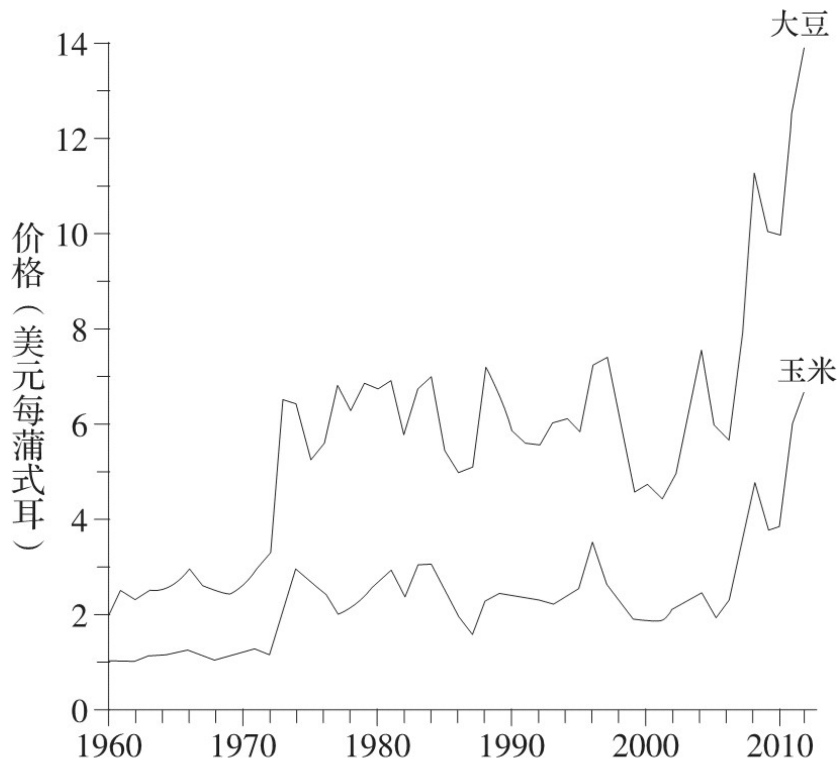


图 17.3　大豆和玉米的价格

图 17.3 表明，1960 年以来，玉米和大豆价格的上升和下降基本是同步的。从 1960 年到 2012 年，每蒲式耳[②]大豆和玉米的平均价格比率是 2.5，因为生产一蒲式耳大豆的时间是生产一蒲式耳玉米的大约 2.5 倍。由于供给和需求的波动，这个比率在不同年份会出现起伏。不过，这种相对于 2.5 的短期波动可以迅速得到校正，因为每当某种作物更加赚钱时，农民就会扩大这种作物的种植面积。

黄金和白银与大豆和玉米没有任何相似之处。我们没有理由认为金银比一定会回到 34 和 38 之间。

我只看数据

宽客们常常不会考虑他们发现的模式是否合理。他们会说："我只看数据。"实际上，许多宽客拥有物理学或数学博士学位，但是只有最基本的经济和金融知识。不过，他们并没有因此而气馁。相反，他们的无知使他们有勇气在最没有希望的地方寻找模式。

有一个笑话。两位金融教授在人行道上看到了一张一百元的钞票。当一位教授伸手去捡钞票时，另一个人说："别理它；如果它是真的，那么它早就被人捡走了。"金融教授喜欢说，金融市场不会让百元钞票躺在人行道上；也就是说，如果有一种轻松的赚钱方式，那么它早就被人发现了。

这种观点并不是完全正确的。有时，股票和债券存在价格错位。在投机浪潮和金融危机期间，金融市场的人行道上散落着装满百元大钞的手提箱。不过，当你认为自己发现了价格错位时，你应该考虑这种错位有没有一个合理的解释。如果黄金价格相对于白银价格出现上涨，市场是否正在将百元大钞愚蠢地扔在路边？是否存在一个合理的解释？

找到金价和银价的历史相关性并不比找到婚姻和啤酒消费量的历史相关性更加令人信服，它只能说明我们为了寻找相关性花费了很长时间。

在推土机面前捡硬币

20 世纪 80 年代，约翰·梅里韦瑟（John Meriwether）在所罗门兄弟公司确立了自己的名声和财富。他利用债券市场的不完善，寻找利率的小幅异常，并且押宝这些异常将会消失。他买入利率过高的债券，卖出利率过低的债券，然后等待利率收敛。

离开所罗门公司以后，梅里韦瑟在 1994 年创立了长期资本管理公司。公司的管理团队包括多位曾在所罗门套利小组工作过的麻省理工

博士、两位即将在1997年获得诺贝尔经济学奖的金融教授（期权专家迈伦·斯科尔斯和罗伯特·C.默顿）以及另一位麻省理工博士戴维·马林斯（David Mullins）。马林斯曾经担任联邦储备委员会副主席，他本来有机会接替艾伦·格林斯潘成为美联储主席。

凭借这个星光闪耀的团队，公司毫不费力地吸引到了许多投资者，尽管其最低投资额是100万美元，而且公司只向投资者透露了一个策略：他们收取的管理费用包括资产的2%以及25%的利润。一位著名的资金经理告诉我，他没有投资于这家公司，因为他唯一确信的事情就是这家公司很贪婪。其他投资者则没有那么谨慎。长期资本管理公司筹集了超过10亿美元资金。

公司的早期策略之一与到期日略有差异的美国国债有关，比如刚刚发行的30年国债和几个月以前发行的30年国债。二者的利率应当是基本一致的，但新发行的债券往往具有更加活跃的交易，因此短期进出市场的交易者更喜欢这种债券。债券市场为之前发行的债券制定稍微高一些的利率，因为这种债券寻找买家比较困难。

公司的收敛赌注是，当新的30年债券发行时，之前的债券之间的差异就会基本消失，它们的利率就会收敛。因此，公司买入早期债券，卖出最新债券，等待它们的利率收敛。斯科尔斯没有使用在人行道上拾起百元钞票的比喻；相反，他将公司描述成巨大的真空吸尘器，认为它正在吸走被其他人忽视的硬币。

凭借25比1的杠杆，长期资本管理公司的真空吸尘器1994年赚了28%，1995年赚了59%，1996年赚了57%，1997年至少赚了22%（扣除管理费之前）。这种巨大杠杆的问题是，即使预想中的收敛暂时转为发散，公司也会受到致命打击。

下面这个假想的简单例子可以说明他们的策略和风险。假设你买入了一份“10年后以20美元购买一瓶葡萄酒”的期货合约。你还卖出了一份“10年后以21美元销售一瓶葡萄酒”的期货合约。如果两瓶葡萄酒是相同的，说明市场效率不高，因为你可以在交割日用20美元买到葡萄酒，同时用21美元将其卖出去，从而稳赚1美元利润。

如果葡萄酒是不同的，这种利润就无法保证了。如果你买入的是赤霞珠葡萄酒，卖出的是雷司令葡萄酒，那么你在交割日不一定能够用一种葡萄酒换到另一种葡萄酒。也许，你发现这两种葡萄酒通常具有相同的价格，希望它们在交割日也能具有相同的价格。你的计划是在交割日用 20 美元购买赤霞珠，以当天的市场价将其卖掉，用得到的钱购买雷司令，然后以 21 美元将其卖掉。表 17.1 说明，如果不出意外，你将赚到 1 美元利润。不管赤霞珠和雷司令在交割日的价格是 10 美元还是 30 美元，只要它们价格相同就可以了。

表 17.1　盈利的葡萄酒收敛

	两种葡萄酒的价格均为 10 美元	**两种葡萄酒的价格均为 30 美元**
买入赤霞珠	-20 美元	-20 美元
卖出赤霞珠	10 美元	30 美元
买入雷司令	-10 美元	-30 美元
卖出雷司令	21 美元	21 美元
净利润	1 美元	1 美元

只要两种葡萄酒的价格在交割日收敛，你就可以获得利润。不过，正如英国一句古老的谚语所说，“即使你把杯子举到嘴唇跟前，你也有可能遇到许多不测。”赤霞珠和雷司令过去的价格接近，但这并不能保证它们未来价格接近。

而且，即使二者的价格真的在交割日收敛，你也只能在那时获得利润。毕竟，如果两种葡萄酒今天的价格不同，我们没有理由认为这种价格差异不会在明天变得更大。在巨大的杠杆作用下，这两种风险促成了长期资本管理公司的覆灭。

长期资本管理公司对历史上具有相关性的债券、股票和抵押贷款进行了复杂的交易。其他交易涉及期权和其他一些证券，它们的价格取决于人们感受到的风险。当其他投资者认为风险变大时，资产价格往往会发生变化，此时公司就会下注，因为公司觉得人们感受到的风险很快就会回到历史上的正常水平。

在每一种情形中，公司的策略都是买入相对便宜的资产，卖出相对昂贵的资产，利用其中的落差赚取利润。公司名称中的"长期"一词显然表明，长期来看，价格将会收敛，利润将会浮现出来。不过，正如伟大的英国经济学家约翰·梅纳德·凯恩斯在大萧条中观察到的那样：

> 这种长期趋势是对当前事务的误导。长期来看，我们都会死去。如果经济学家在混乱时期只能告诉我们"风暴过后海洋还会恢复平静"，这说明他们为自己布置了一项过于简单、毫无意义的任务。

凯恩斯嘲笑的是这样一种观念：长期来看，经济将会平静下来，每个想要找工作的人都能找到工作。凯恩斯相信，经济衰退的风暴比没有人能够活着看到的长期假设更加重要。

现在来看，长期资本管理公司当时是在推土机面前盲目地捡硬币——收集微小的利润，忽视了灾难性风暴的可能性。公司的一个错误是相信历史数据是对未来的可靠指引。公司的一位经理指出："我们所做的事情是依赖经验。我们所有的科学都是建立在经验的基础上。如果你不愿意根据经验得出任何结论，那么你最好把双手放到背后，什么也不做。"

无所事事总比做一些愚蠢的事情要好。不去交易黄金和白银好于相信金银比总会回到 34 和 38 之间。在长期资本管理公司的例子中，无所事事好于做出"未来将会与过去类似"的假设。

公司的一部分交易具有合理的基础，比如刚刚发行的和最近发行的国债利率应当基本相同。其他一些交易所依据的历史模式并不比金银比更加可靠。例如，公司在英国、法国和德国的利率关系上投入了很大的赌注。全球利润常常朝着同一个方向变化，但我们没有理由认为它们应该同步变化。

全球利率之所以具有相关性，是因为投资者可以对他们购买的债券进行选择。如果一个国家的国债利率低于相似国家发行的国债，那么没有人会购买前一个国家的国债。所以，两个国家不得不制定相同的利率。这很有道理，如果两个国家相同的话。不过，现实中的国家是不同的，它们的债券也是不同的。投资者可能担心某个国家国债违约的可能性比较大。投资者可能担心某个国家的货币贬值，使其国债的实际回报率变低。投资者可能担心某个国家改变税收法律，或者某个国家的债券市场缺乏流动性，使他们很难找到买家。

图 17.4 显示了 1960 年到 1995 年的 35 年时间里法国和德国政府发行的 10 年期债券的利率。两种利率存在紧密的相关性，除了它们不紧密相关的时候。如果你看到法国和德国的利率在 1975 年分离，然后预测二者很快就会再次重合，那么你需要等待 15 年的时间。这非常类似于预测赤霞珠和雷司令葡萄酒的价格将会收敛，然后等待这种预测变成现实的做法。这不是没有风险的赌注。

长期资本管理公司还相信美国抵押贷款利率和国债利率之间存在紧密的关系，但是二者也不是同步变化的。二者的违约风险差别很大，而且抵押贷款拥有更大的复杂性，因为房主可以在搬家时或者以更低的利率获得资金时提前付清抵押贷款，这些混杂因素随时间变化，而且具有众所周知的预测难度。一位经理后来惋惜道："我们有一些专业学者，他们进入公司时没有任何交易经验，但他们立即开始设计模型。根据他们制定的假设，他们的交易看上去也许不错，但是这些假设常常无法通过最简单的常识性检验。"

第二个错误是认为不同赌注（比如对意大利债券、德国股票和美国押抵贷款的赌注）在很大程度上是不相关的，因为它们在历史上是

不相关的。实际上，公司忽视了这些赌注的两个共同点。

长期资本管理公司 1998 年年初的净值接近 50 亿美元。当年 8 月，发生了一场预料之外的风暴。俄罗斯债务出现违约，各个金融市场的风险预测值都出现了上升。长期资本管理公司在许多不同的市场上押下了赌注，但是其中的许多赌注都在押宝风险溢价下降。俄罗斯违约后，历史上毫不相关的投资突然具有了高度相关性。各个领域的风险溢价都在上升，推土机碾过了硬币收集者。

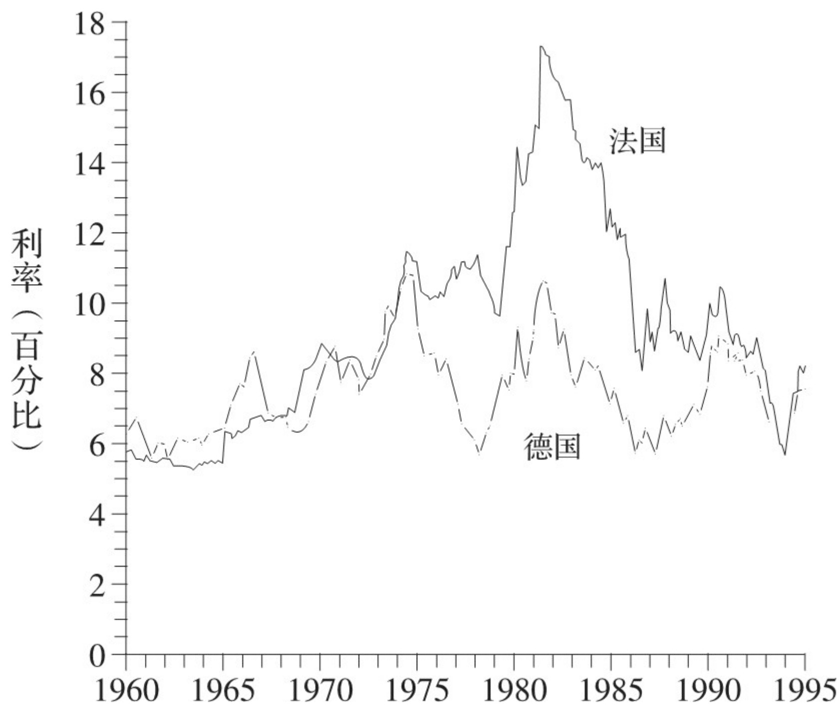


图 17.4　法国和德国 10 年期政府债券的利率

此外，在长期资本管理公司的鼓舞下，许多山寨公司要么猜出了长期资本管理公司的动作，押下了相同的赌注，要么使用类似的统计模型，得到了相同的赌注。当这些赌注出问题时，许多山寨公司开始平仓，卖出长期资本管理公司买入的资产，买入长期资本管理公司卖出的资产。价格没有收敛，反而越来越发散。长期资本管理公司陷入了很大的麻烦。

长期资本管理公司声称，它所需要的仅仅是时间而已。只要等待足够长的时间，金融市场就会恢复正常。不过，他们已经没有时间了。他们巨大的杠杆带来了巨大的损失。公司 8 月 21 日亏损了 5.50 亿美元，整个 8 月亏损了 21 亿美元，占公司净值的将近一半。公司试图筹集更多资金，以度过这场风暴，但是受到惊吓的投资者已经不想再向公司贷款了。他们想把自己的钱收回来。

凯恩斯不仅是一位经济学大师，也是一位具有传奇色彩的投资者。他的一句广为传诵的评论是："市场维持非理性的时间可能会超出你拥有偿付能力的时间。"也许，市场对俄罗斯的违约反应过度。也许，长期资本管理公司的亏损最终会转变成盈利。不过，公司的偿付能力已经支撑不到那一天了。

9 月 23 日，沃伦·巴菲特给长期资本管理公司发送了一份一页纸的传真，提出以 2.50 亿美元收购公司的建议，这相当于公司年初净值的大约 5%。这份提议不接受讨价还价，而且将在中午 12:30 过期，这与传真的发送时间只隔了一个小时左右。截止时间很快过去了，公司没有抓住机会，葬礼的准备开始了。

纽约联邦储备银行担心长期资本管理公司违约的多米诺骨牌效应将会引发全球金融危机。美联储和长期资本管理公司的债权人接管了公司，并且投入了足够多的资金，以争取时间对公司的资产进行有序清算。债权人收回了自己的资金，公司创始合伙人损失了 19 亿美元，其他投资者获得了一个昂贵的教训：杠杆具有巨大的威力，统计关系是靠不住的。

长期资本管理公司破产以后，梅里韦瑟和公司的其他许多合作伙伴根据几乎相同的策略创办了新的基金。大多数基金公司在 2007 ~ 2009 年的金融危机期间破产，其原因往往与长期资本管理公司失败的原因相同。没关系。梅里韦瑟迅速创办了另一家基金公司。

骗我一次是你的耻辱，骗我两次是我的耻辱。骗我三次……你懂的。

闪电崩盘

优秀科技杂志《连线》2011 年的一篇文章对计算机股票交易系统充满了敬畏和赞美之情。这种系统被称为“算法交易员”，因为计算机可以根据计算机代码（“算法”）代替人类决定买入和卖出。指导计算机的算法是由人类编写的，但是在那以后，计算机将自行决定交易。

一些系统跟踪股票价格，其他系统对新闻报道进行剖析。它们都在寻找过去曾经作为盈利信号出现的模式。当这种模式重新出现时，它们就会迅速行动，在几秒钟的时间里买入成千上万的股票，然后通常在几秒钟以后将其卖出。它们不断进行这种操作，日复一日，成千上万的股票几秒钟之内几个美分（甚至几分之一美分）的利润可以累积成可观的财富。《连线》赞美道，这些自动化系统“比人类更加高效，更加迅速，更加聪明”。

这些程序对数据的处理的确比人类更加迅速，但它们并不比编写代码指导计算机的人类更加聪明。如果人类让计算机寻找潜在盈利模式——不管它们发现的模式是否合理——并在这些模式重新出现时买入或卖出，计算机就会执行这种命令——不管它是否合理。实际上，一些编写计算机程序的人吹嘘说，他们其实并不理解为什么计算机做出这样或那样的交易决定。毕竟，计算机比他们更加聪明，不是吗?他们应该做的不是自夸，而是祈祷。

这与长期资本管理公司使用的数据挖掘方法类似，但它的问题更加严重，因为人类的所有判断都被取消了。而且，模仿问题被放大了 100 倍。如果人类为数百台计算机提供非常相似的指导，那么数百台计算机可能会试图同时买入或卖出相同的事物，使金融市场出现极大的动荡。幸好，《连线》杂志认识到了无人监管的计算机同步行动的危险：“在最糟糕的情况下，这将是一种难以理解的反馈回路……可能会将它们所依赖的系统压垮。”

2010 年 5 月 6 日，美国股票市场遭遇了所谓的“闪电崩盘”。当天的投资者对希腊债务危机感到紧张，一位焦急的共同基金经理试图销售 41 亿美元的期货合约，以对冲其投资组合的风险。他的想法是，如果市场下跌，基金中的股票组合资产的损失可以被期货合约的利润抵销。不知为什么，这种看上去比较谨慎的交易触发了计算机的操作。计算机买入了该基金销售的许多期货合约，并在几秒钟之后将其卖出，因为它们不喜欢长期持仓。期货价格开始下跌，计算机决定更加密集地买入和卖出。这些计算机陷入了一种交易狂热之中，它们相互买入和卖出期货合约，就像传递烫手的土豆一样。

没有人知道计算机被触发的准确原因。还记得吗？即使是编写计算机程序的人也不理解它们的交易。在 15 秒之内，计算机相互之间交易了 2.7 万份合约，占交易总量的一半。在 15 秒的狂热结束时，净买入合约只有 200 份。这场交易狂热传播到了常规股票市场，沽盘订单像洪水一样淹没了潜在购买者。道琼斯工业平均指数在五分钟之内下跌了将近 600 点。市场价格乱了套，但计算机仍然在进行交易。坚实的蓝筹公司宝洁的股价在不到四分钟的时间里下跌了 37%。一些计算机以每股超过 10 万美元的价格买入苹果、惠普和苏富比的股票。其他计算机以每股不到一美分的价格卖出埃森哲和其他大公司的股票。计算机并没有常识。它们盲目地买入和卖出，因为这就是它们的算法要求它们去做的事情。

这场狂热之所以结束，是因为期货市场的内置保护机制将所有交易中止了 5 秒钟。令人难以置信的是，5 秒钟的价格稳定足以说服计算机停止疯狂的交易。15 分钟后，计算机恢复了正常，道指暂时性的 600 点下跌仅仅给人们留下了一个荒诞的回忆。

此后还发生了其他一些闪电崩盘，未来很可能还会发生更多类似的事件。巧合的是，2013 年 8 月 30 日，宝洁在纽约股票交易所（NYSE）再次中招。这是一次迷你闪电崩盘，因为纽交所的其他股票并没有受到影响，宝洁在其他交易所的股票也没有受到影响。

由于无法解释的原因，宝洁在纽交所的股票在一秒钟之内发生了200次交易，一共涉及大约25万股股票，导致股价下跌5%，从77.50美元降至73.61美元，然后在不到一分钟以后恢复正常。一个幸运的人恰好在正确的时间和正确的地点购买了6.5万股股票，迅速获得了15.5万美元利润。为什么会发生这样的事情？没有人知道。别忘了，人类并不比计算机更加聪明。

是的，没错。

×如何轻松识破一本正经地胡说八道

> 不要把资金押在历史模式以及几乎没有合理解释的关系上。黄金价格也许是白银价格的34到38倍，但是它没有理由必须维持在这个范围内。美国押抵贷款市场也许多年来与俄罗斯债券市场没有相关性，但是一场恐慌可能对两个市场产生类似的影响。不要仅仅观察数字。应该考虑原因。

① 1盎司约为31.1克——译者注

② 1蒲式耳约为36.4升——译者注

第 18 章
增长的极限

本书的许多例子涉及一种狡猾的做法，那就是在没有基本理论或明确目的的情况下搜刮数据，而不是发现可能带来名声和资助的有意义的统计关系。这属于缺乏理论的数据。

另一个极端是缺乏数据的理论——将半真半假的理论作为事实提出，但却从不用数据对其进行检验。如果一种理论没有得到可靠数据的检验，那么它仅仅是一种猜测。对于未来几十年甚至几百年的预测来说，这种现象尤其明显。

穷途末路

当阿尔伯特·爱因斯坦被问及他所了解的最重要的概念时，他立即给出了“复利”这个答案。不是统计力学。不是量子理论。不是相对论。是复利。

假设你投资 1000 美元，年利率是 10%。第一年，你的 1000 美元投资赚到了 100 美元利息。此后每年，你都会得到 1000 美元的 10%，外加你已经赚到的利息的利息，这就是使“复利”强大到近乎神奇的原因。经过 50 年的利滚利，你的 1000 美元将会变成 117391 美

元。经过多次复合计算，一个看上去比较温和的回报率可以将一笔小额投资转变成一笔财富。复利的奇迹不仅仅适用于投资。它适用于一切以复合速率增长的事物，包括人口和资源消耗量。

1972 年的报告《增长的极限》很好地展示了复利的奇迹。这篇报告是由听上去很神秘的“罗马俱乐部”赞助的。罗马俱乐部是神秘的兄弟会吗？它是否与共济会或《达·芬奇密码》有关？实际上，罗马俱乐部是一个很普通的组织。1968 年，一位意大利实业家在罗马成立了一个跨国组织，以支持对世界前景的长期跨学科分析。真是一项雄心勃勃的事业！实业家将这个具有宏大目标的小型组织称为罗马俱乐部。

《增长的极限》很受欢迎。它被翻译成 30 种语言，销量超过了 1200 万本。首席作者都尼勒·“达纳”·梅多斯（Donella “Dana” Meadows）拥有化学学士学位和生物物理学博士学位，曾经参与杰伊·W.福雷斯特（Jay W. Forrester）组织的一个麻省理工团队，团队的任务是制作世界社会经济系统模型“世界 3”。

福雷斯特在职业生涯早期是一位电气工程师，后来从工程领域转到了管理领域，提出了用于理解复杂现象的所谓“系统动力学”。为了理解一个系统的动态演化（比如一个新产品的创造、生产和营销），我们应该理解系统各个部分的反馈、相互作用和延迟。由于这些模型非常复杂，因此需要通过计算机模拟来分析系统的演化。

福雷斯特受邀参加了罗马俱乐部的一次会议。会上，人们问他能否用系统动力学模拟不断增长的人口需求和有限的地球资源导致的“人类困境”。答案当然是肯定的。虽然福雷斯特没有接受过正式的经济学培训，但他在返回美国的飞机上设计出了“世界 1”模型。他对模型进行了调整，得到了涉及 43 个变量的“世界 2”，并在 1971 年的《世界动力学》一书中对这个模型的一些计算机仿真进行了描述。

福雷斯特的结论是，世界 1971 年的生活标准也许是这个星球能够承载的最高水平。他总结道：“从今后 100 年的长期视角来看，发展中国家目前的工业化努力也许是不明智的。”为了维持 1971 年的生活标

准，他建议将出生率降低 30%，将粮食产量降低 20%，将自然资源使用量降低 75%。

同“世界 2”相比，梅多斯团队开发的“世界 3”模型添加了更多的变量和方程，但它具有相同的核心假设和悲观结论。它也具有相同的缺陷。最根本的问题是，福雷斯特和梅多斯不是经济学家，而且似乎对经济理论和数据一窍不通。没有一个方程以经济增长模型为基础；没有一个方程来自对历史数据的估计。

对未来几十年或几百年进行推测是一项极具挑战性的任务。在这个例子中，他们试图在不使用经过其他人检验的理论和不对他们自己的理论进行检验的情况下完成这项任务。他们的模型是“输入垃圾，输出福音”的一个极端案例，因为一些人之所以相信这些模型的预测，不是因为这些预测多么合理，而是因为这些预测来自计算机。

作者在《增长的极限》20 周年更新版中做出的推荐很好地说明了这种模型与现实之间的脱节：

> （这个场景）显示了一个以“足够”为定义的模拟世界……这个世界决定以每人每年平均工业产值 350 美元为目标——大约相当于 1990 年韩国的产值或者巴西产值的两倍……如果这个假想的社会能够同时减少军事开支和腐败，那么人均工业产值为 350 美元的稳定经济体在物质享受上相当于 1990 年欧洲的平均水平。

耶鲁大学著名经济学家威廉·诺德豪斯（William Nordhaus）是这些模型的坚定批评者。在一篇题为《致命模型 2》的论文中，他对这种模型（他称之为“限制二”）做出了下列嘲讽：

> 这段推荐文字的事实性论断错误百出，你甚至怀疑“限制二”是否正在谈论另一个星球。1990 年全球人均国民生产总

值的粗略估计值是 4200 美元，经济合作与发展组织成员国则是 20170 美元。按购买力平价计算，韩国 1990 年的人均国内生产总值是 7190 美元，不是 350 美元……“限制二”的建议将会把我们的物质追求限制在索马里或乍得的生活水平上……世界将没有能力投资于延缓全球变暖或研发节能科技的事业中。这种（限制增长的）处方将以牺牲地球居民为代价挽救这个星球。

福里斯特和梅多斯构造的模型涉及许多变量和方程的复杂相互作用，但他们做出的灾难性预测的核心非常简单：复合增长将以惊人的速度耗尽有限的资源。

“世界 2”模型假设世界上的资源是世界当前每年资源使用量的 400 倍。如果使用量维持恒定，这些资源可以维持 400 年。不过，如果使用量每年随着人口的增长而增长，那么资源的枯竭时间将会提前。复合增长既具有成就能力，也具有毁灭能力。如图 18.1 所示，按照 2% 的温和增长率计算，100 年后的年度使用量将达到当前水平的七倍，世界上的资源将在 110 年以后耗尽。

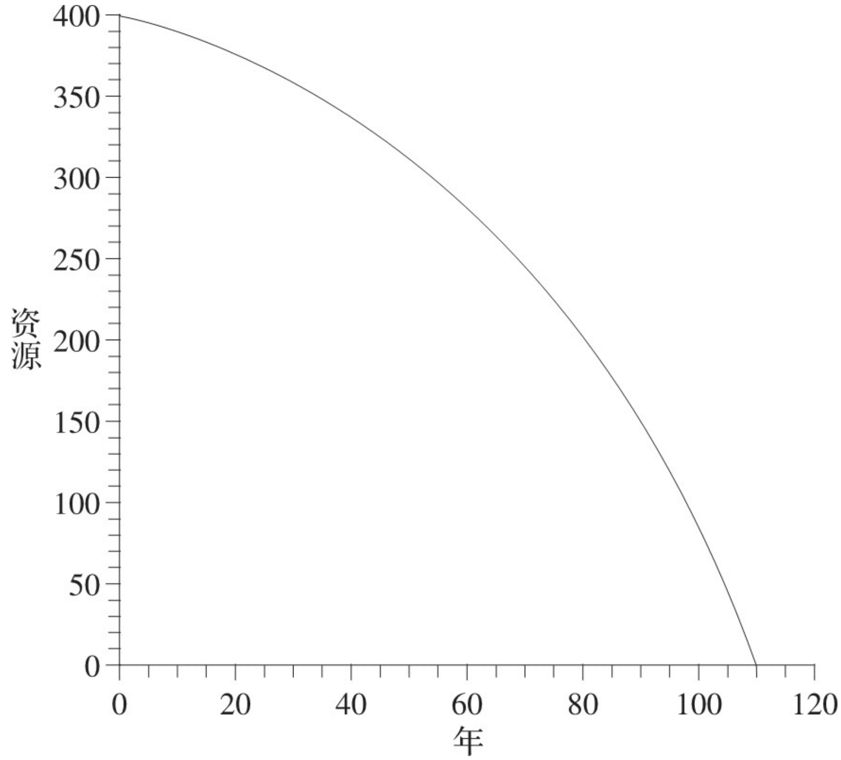


图 18.1　世界末日是不可避免的

复合增长模型对假定资源水平和使用量的增长速度并不敏感，因为不管资源的初始水平如何，不管使用量的年增长率如何，有限的资源最终都会耗尽。真正值得注意的是，在复合增长的作用下，世界资源的耗尽时间极为短暂。根据 1971 年发表的“世界 2”模型，世界的生活标准将在 1990 年达到峰值，然后不可逆转地下降。

解决方案呢？福雷斯特的模型认为出生率与整体生活标准存在正相关，尤其与粮食产量存在正相关。他相信粮食产量的增长将会导致出生率增长，因此他建议减少粮食产量！具体地说，他认为，各国政府应该将全球食品供应量减少 20%，以说服人们生养更少的孩子，否则他们就会挨饿。（你可能认为他在开玩笑，但他并没有开玩笑。）

世界末日式的人口预测至少在 1798 年就出现了，因为马尔萨斯在这一年出版了《人口论》。当时的许多知识分子认为，人类社会已经进入了可以无限发展的阶段。马尔萨斯的观点和这种无拘无束的乐观

主义相反，他指出，在历史上，人口的增长速度往往高于食品供应的增长速度。实际上，农业生产力的提高只会导致更高的出生率，直到人口过多，食物不足为止。这种危机是通过饥荒、疾病和战争解决的——这掩盖了最根本的问题，即人口过剩。因此，周而复始的循环一直在继续。由于马尔萨斯冷酷的分析，经济学有时被贴上“悲观科学”的标签。

马尔萨斯认为，要想解决人口过多、食品不足的问题，他们可以不必提高死亡率，而是通过独身、节育、堕胎和卖淫降低出生率。是的，这位优秀的教士看到了堕胎和卖淫的好处。

到目前为止，马尔萨斯的观点是错误的。惊人的科技发展使世界人口增长了 7 倍。不过，福雷斯特和梅多斯实际上重复了马尔萨斯 1798 年的假设和结论。资源使用量以复合增长率增长，资源的供应量则不会增长。如果发现新的资源，更高的生活标准将会提高出生率，因此人口的增长很快就会超过资源的增长。

马尔萨斯、福雷斯特和梅多斯所持观点的一个明显问题是，富足并不会提高出生率。自从马尔萨斯那个时代以来，我们发明了有效的节育途径。而且，不管我们比较不同国家还是考察一个国家，生活标准的提高显然会降低人口增长速度。今天，一些发达国家的人口增长率近乎为零，甚至变成了负值。

此外，“资源使用量以复合增长率增长，世界上的资源供应量则是固定的”这一假设没有任何经验基础。能够以更高的效率使用资源的技术进步可能也会以复合增长率增长。马尔萨斯、福雷斯特和梅多斯完全没有注意到，价格会导致人们适应变化的环境。如果某样东西变得稀缺，它就会变得更加昂贵，人们就会减少对它的使用，并且发现不那么昂贵的替代事物。

福雷斯特和梅多斯将所有自然资源打包放在一起，将其称为“资源”。在现实中，有许多不同种类的自然资源，其中一些资源比另一些资源更加丰富（想一想氧气和黄金）。而且，人类具有无穷无尽的智慧，可以想办法用更加丰富的资源替代不那么丰富的资源。我们已经

学会了用核燃料代替化石燃料，用电子邮件代替“蜗牛”邮件，用塑料代替木头、金属和玻璃。没有人知道未来100年将会发生哪些替代，但是我们可以肯定，未来一定会发生新的替代。

马尔萨斯、福雷斯特和梅多斯所持观点的根本问题是，他们的推理具有一丝合理性，但他们并没有考察历史数据，以检验这些数据是否支持他们的理论。他们相信这些理论，所以他们要求我们也相信这些理论。

我们可能被缺乏数据的理论欺骗，正如我们可能被缺乏理论的数据欺骗。

×如何轻松识破一本正经地胡说八道

我们既需要理论，也需要数据。不要仅仅被其中的一种事物说服。

如果有人通过搜刮数据发现了某种模式，我们还需要一种合理的理论。另一方面，在经过令人信服的数据检验之前，理论仅仅是理论而已。

不管一项研究是谁做的，它都需要通过常识性检验，而且需要通过没有被数据挖掘所污染的无偏数据的检验。

第 19 章
何时相信，何时怀疑

如今，我们被数据包围。同时，我们又常常被数据欺骗。

有时，数据可以帮助我们评估相互竞争的观点，做出良好的选择。政府选择能够降低失业率，消除霍乱、小儿麻痹症和天花的政策。企业开发能够改善我们的生活、并能以合理成本制造的产品。投资者购买低成本共同基金，选择由价格具有吸引力的股票组成的分散式投资组合。病人被给予能够真正挽救生命的药物和治疗。消费者选择有效的、价格具有竞争力的产品。我们停止吸烟，开始锻炼身体。

其他时候，我们被数据欺骗，做出糟糕的选择，导致昂贵甚至灾难性的后果。政府在经济衰退期间实行租金管制，增加税收。企业追逐最新的管理风潮，依赖于具有严重偏差的调查。投资者追逐热门股票，将他们的资金托付给行骗高手。病人对检测结果产生误解，接受毫无价值甚至有害的药物和治疗。消费者追随最新的减肥热潮，并且因为房价一直在上涨而买房。我们认为成功者受到了诅咒，失败者即将时来运转。

有时，我们应该相信数据；有时，我们应该怀疑数据。

被模式诱惑

我们生来倾向于以某种方式理解周围的世界——发现模式，并且编造出解释这些模式的理论。我们低估了毫无理由的随机事件生成幸运或不幸模式的容易程度。

我们应该努力意识到，我们很容易受到模式的诱惑。我们不应该受到引诱，应该保持怀疑的态度。相关、趋势和其他模式本身无法证明任何事情。如果没有合理的解释，任何模式都仅仅是一种模式而已。每一种合理的理论都应该接受新数据的检验。

具有误导性的数据

在实验中，研究人员常常改变一个因素，同时将其他混杂因素维持恒定，然后查看结果。例如，我们可以为植物提供不同的肥料剂量，同时将水、阳光和其他因素维持恒定。不过，在行为科学中，涉及人类的实验是受限的。我们无法让人们辞去工作，与配偶离婚，或者生孩子，以便观察他们的反应。相反，我们只能使用观测性数据——观察失去工作、离异或者有孩子的人。我们很容易根据观测到的现象得出结论。我们都在做这样的事情，但它是有风险的。

当我们比较做出不同选择的人，同时不去考虑他们为什么做出这些选择时，就会出现自选择偏差。一些学生之所以选择毕业率较低的大学，可能是因为他们不太可能从其他大学毕业。选择上大学的学生可能比选择不上大学的学生更聪明，更有动力。

如果消除选择因素，我们的结论可能更有说服力；比如学生被随机分配到不同的大学或者被随机推向社会。幸运的是，科学家不能为了收集数据发表研究论文而毁掉人们的生活。遗憾的是，我们需要对那些可能存在自选择偏差的数据保持警惕。

为我们观察到的现象寻找理论的另一个常见问题是幸存者偏差，因为我们无法看到不复存在的事物。对于老年人的研究不会包含中年

夭折的人。对于结束轰炸任务的飞机的考察不会包含被击落的飞机。对于入住某些酒店、搭乘某些航班或者访问某些国家的人群的调查不会包含有过一次经历以后再也不来的人。

对于伟大公司共同特点的整理不会包含拥有这些特点但不那么成功甚至已经破产的公司。如果我们在观察数据之前列举出我们认为重要的因素，并且找出拥有这些特点和没有这些特点的公司，然后再去查看实际结果，我们的结论将会更有说服力。

不要忽视记录数据或编写计算机代码时出现差错的可能性。为了支持“在大萧条中减少政府开支，增加税收”的观点，莱因哈特和罗戈夫无意中忽略了与他们的观点相冲突的数据，并且以一种不同寻常的、无法令人信服的方式计算平均值。莱维特和多诺霍认为美国的合法堕胎降低了犯罪率，但他们的研究包含许多错误，包括一个程序错误。

变形的图像

作为视觉信息，图像可以帮助我们解读数据，做出推断。有用的图像可以准确而一致地显示数据，帮助我们看到倾向、模式、趋势和关系。一张图片的价值抵得上一千个数字。

不过，图像也会歪曲数据。当心那些通过忽略数轴零点夸大差异的图像。还记得吗？当数轴没有显示零点的时候，一位总裁曾经认为收入的温和下降是一场危机。

当图像拥有两个纵轴而且其中的一个或两个纵轴忽略零点时，应当格外小心。聪明的艺术家可以对数轴上显示的数据范围进行调整，使收入的增速看上去高于消费品价格，或者消费品价格的增速看上去高于收入。

当心那些省略数据、使用不一致的数轴间隔、颠倒数轴、塞满图表垃圾的图像。

缺乏思考的计算

人们有一种自然而然的倾向，那就是仅仅关注数值计算的准确性，不去深入思考这些计算是否正确。

在著名的蒙提·霍尔问题中，好和坏两种结果的存在性并不意味着它们的可能性是相等的。更加令人吃惊的是，专家们不断抛弃常识，认为没有出现的孩子的性别取决于这个孩子是否比同胞大，是否拥有不同寻常的名字，或者是否出生在星期二。在你核对某人的计算之前，应首先核对他的推理。

一个很常见的逻辑错误是混淆两种条件性陈述。患病的人得到阳性检测结果的概率并不等于得到阳性检测结果的人患有疾病的概率。

人口随时间增长，许多人类活动也是如此，包括看电视的人数、吃橘子的人数以及死亡人数。这些数据是不相关的，但它们存在统计相关性，因为它们都会随着人口的增长而增长。看电视并不会导致我们吃桔子，吃桔子也不会导致死亡。在统计学中，相关性并不是因果关系的代名词。不管两种事物的关系多么紧密，在做出判断之前，我们都需要一种合理的解释。

在实证研究中，比较常常是非常重要的。不过，我们需要当心肤浅的比较：对于大数和小数百分比变化的比较，对于除了随时间增长以外没有任何共同点的事物的比较，以及对于不相关数据的比较。

寻找混杂因素

当你听到某种令人困惑的（甚至合理的）论断时，应当考虑是否存在混杂因素的作用。瑞典的女性死亡率高于哥斯达黎加——因为瑞典的老年女性比较多。伯克利研究生计划录取的女性申请者比较少——因为女性申请了难度比较大的计划。一种手术比另一种手术更加成功——因为它被用于更加轻松的案例。患有胰腺癌的病人比其他病人饮用更多的咖啡——因为其他许多病人存在溃疡，戒掉了咖啡。

混杂因素常常存在于观测性研究中，此时我们无法控制人们的选择；但它也存在于实验环境中，因为研究人员有时会忘记对某个混杂因素进行控制。

手气好

在抛掷十次硬币时，连续出现四个正面的结果看上去很奇特，似乎无法用偶然性来解释。实际上，连续四个及以上相同结果出现的概率是 47%。不过，我们很容易认为硬币或者抛硬币的人一定存在某种独特之处。

当篮球选手连续多次投篮命中时，我们很容易认为选手的手气很好——手气好的选手下次投篮很可能也会命中。我们很容易低估这种现象仅仅出于巧合的可能性。同样的道理适用于所有体育项目以及体育以外的领域。

一些可信的证据表明，运动员有时的确会进入火热的状态，但他们的能力波动比选手和球迷想象的要小得多。同样的道理无疑也适用于体育以外的领域。

均值回归

在一次测试中得到最高分的学生（让我们叫她雷切尔吧）很可能受到了好运的帮助（被问到熟悉的问题，猜出正确的答案），因此她的表现不仅优于其他学生，而且优于自己的能力。雷切尔的能力与平均能力的差距很可能不像她的测试分数与平均分数之间的差距那么大。因此，她在第二次测试中的分数很可能更加接近平均分数——也就是向均值回归。类似的，得到最低分的学生的能力很可能不像测试分数表现得那样可怜；这名学生很可能会在第二次测试中做出更好的表现。

最成功和最不成功的公司、运动员、员工、雇主和潜在灵魂伴侣也是如此。不要被成功和失败欺骗。最优秀的人与平均水平的差距很可能不像看上去那么大。最糟糕的人也是同样的道理。位于极端位置的事物很可能会向均值回归。

平均定律

一种非常不同（同时也极度不正确）的观点是，成功一定会得到失败的平衡（反之亦然），以便使事物整体上呈现平均水平。每次硬币正面朝上都会提高背面朝上的可能性。轮盘赌的每个红色结果都会提高黑色结果的可能性。多次击出全垒打的棒球选手即将出局。从未失火的房屋即将被烧毁。经历多次安全飞行的旅客即将遭遇坠机事故。这些信念都是不正确的。

好运当然不会永远持续，但是不要认为好运会提高噩运的可能性，反之亦然。

德克萨斯神枪手

追求名声和资助的研究人员常常会变成德克萨斯神枪手，他们随机开枪，并在弹孔最多的区域绘制靶心。如果你用某种数据编造理论，那么你很容易发现这种理论与数据相符。这样的结果只能证明某人对数据聚集或其他模式进行了寻找。只有当这种结论言之有理，并且得到未经污染的数据检验时，它才是令人信服的。

其他一些德克萨斯神枪手向几百个目标开火，然后只报告他们击中的目标。他们对几百种理论进行检验，然后只报告最符合数据的理论。他们考察癌症受害者的几百种特点，然后只报告最常见的特点。如果我们最终碰上了一种能够被数据支持的理论，这只能证明我们测试了许多理论。这种理论必须言之有理，而且需要得到新数据的证实。

有时，研究人员会隐藏自己的德克萨斯神枪手身份，我们需要使用一些侦探技巧发现他们的恶作剧。寻找不自然的数据分组。当研究人员似乎只报告了一部分统计检验时，应当保持警惕。不要轻易相信那些与数据相符、与常识不符的理论。当某种理论来自对数据的搜刮时，我们无法用这些数据对理论进行公平的检验。

当心经过修剪的数据

当心那些忽略一部分数据的研究，尤其是当你怀疑这些数据之所以被丢弃，是因为它们不支持研究结果时。一些线索包括数据开始或结束于不同寻常的日期（“出生于 1875 年到 1930 年的篮球选手”），或者被分成不同寻常的类别（“75 岁以上的女性”，“政府债务超过国内生产总值 90% 的年份”）。

这些研究真正证明的结论是，一些追求升迁、终身职位或经费的人可以丢弃不支持荒谬理论的数据，从而找到支持这些荒谬理论的证据。

缺乏理论的数据仅仅是数据而已

人们总是满怀希望地研究数据，寻找跑赢大盘或中彩票的方法，然后得出一些可笑的理论，比如超级碗理论或者让你的朋友玛丽替你购买彩票。不要相信这些胡言乱语。

如果我们足够努力，即使面对随机生成的数据，我们也可以找到某种模式。不管这种模式多么明显，我们都需要一种合理的理论来解释这种模式。否则，我们找到的仅仅是巧合而已。

如果某种理论不合理，应当保持怀疑的态度。如果某种统计结论看上去令人难以置信，不要相信它。如果你对其中的数据和检验进行检查，你通常可以发现一个严重的问题，将结论推翻。

缺乏数据的理论仅仅是理论而已

人们相信令人舒服的事物——因为它们令人舒服。相反的证据被人忽视或丢弃。同承认自己的错误相比，对碍眼的事实视而不见是一件更加容易的事情。为了理解世界，即使是最聪明的人也会变得固执而愚蠢。

人们既可以被缺乏理论的数据欺骗，也可以被缺乏数据的理论欺骗。我们既需要理论，也需要数据。仅仅通过搜刮数据寻找模式是不够的。模式需要得到理论的解释，这种理论应该言之有理，而且需要得到新数据的检验。另一方面，在经过可靠数据的检验之前，理论仅仅是一种猜测。

美好的出生日

情人节（圣瓦伦丁节）的起源和它所庆祝的、有时令人费解的爱情一样神秘。传说，一个名叫瓦伦丁（Valentine）的人公然反抗罗马皇帝克劳狄乌斯二世（Claudius II），他为基督教情侣主持婚礼，并且帮助被罗马人迫害的基督教徒。克劳狄乌斯逮捕了瓦伦丁，试图说服他改信罗马异教。瓦伦丁拒绝了他，反而试图说服克劳狄乌斯信仰基督教。他遭到了棍棒和石块的毒打，但他并没有死。于是，他遭到了斩首。

一个名叫圣瓦伦丁（Saint Valentine）的人在 2 月 14 日被埋葬在罗马北部的一条公路旁边。我们并不知道这个圣瓦伦丁是不是与克劳狄乌斯进行过著名对话的那个瓦伦丁，或者这场对话是否真的发生过。

我们也不知道瓦伦丁节是怎样与浪漫的爱情联系在一起的。1392 年，英国诗人杰弗雷·乔叟（Geoffrey Chaucer）写了一首诗，以纪念国王理查二世（Richard II）和波希米亚郡主安妮（Anne of Bohemia）5 月 2 日的订婚纪念日。两个人是在 15 岁结的婚。乔叟写道：“因为这是圣瓦伦丁节，每只鸟儿都来选择它的伴侣。”不过，乔叟指的很可能是 5 月 2 日（纪念热那亚圣瓦伦丁的日子），不是 2 月 14 日，因为英国的 2 月 14 日是冬季，不太可能是鸟儿交配的时节。不过，2 月 14 日已经通过某种方式演变成了全世界销售鲜花、糖果、珠宝和除尘器的节日。

爱情，婚姻，然后是婴儿。孕妇的“预产期”大约是上次正常月经期第一天的四十个星期之后。这是一个平均值，存在很大的变数——就像提前几个星期的天气预报一样。即使考虑到妇女的年龄、种族以及是否是头胎，婴儿在预产期出生的可能性也只有大约 5%。20% 的婴儿出生在预产期两个星期之前或者两个星期之后。

2011 年的一项研究考察了美国 1996 年到 2006 年的数据，认为母亲可以选择在情人节分娩。这是一个令人兴奋的有趣结论，因为自然分娩通常被认为是自发的，妇女无法控制它的时间。

研究发现，妇女不仅可以选择情人节（爱情之日），而且可以避免孩子出生在万圣节（死亡之日）。

且慢。许多出生在万圣节的人在听说这项研究以后感到非常愤怒，他们声称万圣节是自己最喜爱的节日。一个人写道：“我喜欢生日这天的‘不给糖就捣乱’活动。”另一个人写道：“我喜欢生日和万圣节这天的化妆派对。”

民调显示，圣诞节是美国人最喜爱的节日，随后是感恩节和万圣节。情人节和光明节并列第十。由于情人节同罪恶和失望联系在一起，因此它只能获得“最不受欢迎的节日”这一荣誉。

可是，这项研究的的确确是由两位耶鲁教授进行的。

你怎么看？

出版后记

统计学是大数据时代炙手可热的学问。

统计学把乱码一般的数据翻译成意义非凡的逻辑。人们习惯了用数据说话：小到防治疾病、理财投资，大到规划公司经营战略，乃至于制定国家宏观政策，都需要用统计学分析数据，建立模型，以形成决策的依据。

当人们犯了错，常说被数据欺骗了。数据不会说谎，是你会错了意，被统计学欺骗了，形成了错误的认知。在不确定的世界里，最优决策可遇不可求，最不坏的决策才是现实的追求，善用统计学即可达成最不坏的决策。本书就是关于统计学的防骗指南。

著名经济学家罗纳德•科斯曾说："如果你对数据拷打足够长的时间，它一定会招供。"有些人是误用了统计学，得出了错误的结论；有些人——甚至是久负盛名的学者，则是故意用错了统计学，得出了自己想要的结论。无论是何种情况，你都需要擦亮眼睛，不要上当！

本书作者加里•史密斯是美国著名的统计学者，曾在耶鲁教授多年统计学。书中他从统计学的基本原理出发，全面介绍了各种统计误区，犀利地戳穿了久负盛名的《魔鬼经济学》《追求卓越》等超级畅销书的弥天大谎。告诫我们，要对耸人听闻的观点保持警惕，不要轻易否定我们的直觉和常识。根据他所列出的十数条基本统计学原则，你便可以轻松识破生活中常见的误区和骗局，做出最不坏的决策。

诚如作者所言，本书希望你破除对数据的迷信，学会基本的统计学常识，从而具备起码的辨别力。除了本书之外，我公司出版的统计学方面的通俗读物还有《女士品茶》《统计思维》等品种，希望能给您一双看穿复杂世界的慧眼。

服务热线：133-6631-2326　188-1142-1266

读者信息：reader@hinabook.com

后浪出版公司
2017 年 7 月


日産で学んだ 世界で活躍するためのデータ分析の教科書
如何用数据
解决实际问题
后浪
大数据时代必备的数据分析入门指南
让洞见更准确，让沟通更简洁
从订指标、报预算，到分析市场、评估风险，
学会用数据表达想法、说服对方和赢得信赖
数据分析其实比你想象得更简单！
[日] 柏木吉基 著
赵 媛 译
江西人民出版社

如何用数据
解决实际问题

[日]柏木吉基 著；赵媛 译

版权信息

书名：如何用数据解决实际问题
作者：[日]柏木吉基
译者：赵媛
书号：ISBN 978-7-210-10434-6
版权：后浪出版咨询（北京）有限责任公司

目录

序章　我在日产学到了什么

让戈恩社长点头称赞的管理工具

给出“答案”的期限是 3～4 个月

如何在实际工作中使用数据

第 1 章　解决问题，你需要“流程”

第一个重点是“划定范围”

从锁定原因到研究对策

开始之前的准备工作

数据分析需要“假设”

用框架来查缺补漏

开端决定了结论的质量

开始以后很难再扩大视野

思考问题之“外”的问题

「解决问题的故事 1」

问题：汽车销量下滑，怎么办？

第一步：明确计划与现实之间的差距

第 2 章　分解数据，找到“问题的关键”

“趋势”视点和“快照”视点
“WHAT 型假设”
表示大小的“平均值”
“好用”背后的陷阱
平均值不能代表总体
了解“中位数”
“油炸豆腐”和“天妇罗面渣”，哪个更好吃
平均值所掩盖的真相
乌冬面和荞麦面的教训
用“波动”的视点给平均值做补充
标准差为什么“用不上”
“变异系数”的魅力
视觉也是“感觉”
「解决问题的故事 2」
第二步：“分解”数据，锁定问题的关键
第 3 章　采用交叉视点，锁定“原因”
表示二者关系的“相关系数”
Excel 的小妙招
锁定原因也需要“假设”
相关系数的四大优势
找到相关分析的着眼点
奖金被用到哪里了
用 Excel 加载项进行批量分析

利用矩阵排列优先顺序

不要随便编故事

「解决问题的故事 3」

第三步：建立 WHY 型假设，关注影响客户忠诚度的要素

第 4 章　制定对策，要依据“方程式”

10 秒钟完成一元回归分析

关注相关系数的平方

注意事项及应用事例

看相关系数还是看斜率

「解决问题的故事 4」

第四步：通过一元回归分析，发现车型 B 和车型 C 的不同问题

第 5 章　用数据讲故事

解决问题的故事

把解决问题的过程展现出来

在组织中使用数据的价值与难点

你能用数字推翻众人的理解吗

更上一层楼（高级技能简介）

后　记

出版后记

返回总目录

序章
我在日产学到了什么

解决实际业务或者管理中的问题是商务人士每天都要面对的必修功课。或者不妨说，商务人士的工作就是要解决各种各样的问题。其中大概有很多人依靠过去累积的知识或直觉解决眼前的问题。但是，请考虑下面这两个问题：

・以前的经验今后也会畅通无阻吗？

・自己所见的范围和经验没有局限性吗？

关于第一个问题，不难想象，在瞬息万变、新事物层出不穷的当今社会，已经无法仅凭个人或组织的既往经验来判断和解决所有的问题。至于第二个问题，很明显，随着全球化的发展，业务逐渐走向多元化，并不断扩大，从业务的整体来看，个人所能掌握的知识是极其有限的。

我在日产汽车公司的全球总部工作了大约11年。最初的6年，我隶属于掌管海外市场的营销和销售部门，主要工作是通过“数据”来把握所负责国家的销售业绩、当地子公司的经营状况等无法亲自确认的情况，然后根据业绩数据来预测将来，制定预算和计划。

在此期间，日产公司的总部进一步实现了全球化发展，公司里可以看到来自各个国家的面孔，外籍员工的人数与日俱增。在这种情形下，如果部门中资历较老的日本人仅凭自己的直觉或经验发表意见，有时就会被视为过于“主观”，甚至被外籍高管置之不理。

因为是中途转行加入日产公司，我不太会受到过去的条条框框或陈规旧矩的束缚，并且一直有意识地思考，作为半路入行者应该如何为公司创造价值,自己有哪些不同于为公司效力二三十年的老员工的新价值。

我的具体做法是，结合自己所擅长的数字和逻辑，从外部的视角出发，深入挖掘公司及海外市场的数据和信息，反复尝试，直至找出能让自己信服的解释或故事。

让戈恩社长点头称赞的管理工具

当然，并非所有的尝试都能达到预期的效果。不过，在这个过程中，我也发现了一些工作十几年、二十几年的老员工都未曾察觉的问题或事实。我用简单易懂的方式将这些问题展现出来，既实现了自身的价值，赢得对方的信赖，同时也自然而然地为自己找到了立足之地。

高层管理者能否客观地把握全球的业务状况，迅速采取适当的行动，这直接左右着公司的经营。我通过当时的直属上司，将自己开发的管理工具和机制提交给日产 CEO 卡洛斯・戈恩先生。这个工具能够按照相同的指标，对除日、美、欧之外的 120 个国家或地区的销售及经营状况进行客观的数据管理，从而帮助经营者在俯瞰全局的基础上采取适当的措施。而在这之前，一般都是由各地区的负责部门按照各自标准来评价其经营状况，各自采取他们认为必要的措施。

这样不仅无法掌握各地区的判断和措施是否适当，经营者也无法掌握涵盖所有市场的整体情况。因此我决定要深入探究这个问题。

我的想法和思路能够以符合逻辑的数据形式展现出来，因此得到采纳，成为后来的管理工具之一。后来，我成为第一位运用该工具工作的管理者，使它走上常规应用的轨道。通过这段经历，我体会到运用数据来客观地把握和评价现状的重要性及其对公司管理的影响，并认识到逻辑和数字能够在向对方传递信息时发挥重要作用。这为我解决各种问题奠定了基础。

后来，作为公司的内部咨询师，我在专门为戈恩社长以及各部门高层管理者解决管理和实际业务问题的团队度过了 4 年时光。不同于外聘咨询师，我们肩负着参与公司内部政治、提出方案并推动执行、与相关部门构建人际关系等重要责任，是非常值得挑战的岗位。

给出“答案”的期限是 3～4 个月

无论哪个部门的问题，一般都必须在 3～4 个月的短时间内给出答案。大多数情况下，我都需要与相关部门的关键人物一起，推动团队得出结论。

这些结论当然是要呈报给 CEO 和（以外国人为主的）高层管理者的。对有些领域（例如商品企划）的问题，我并没有直接从事相关业务的实际经验，但仍然需要在期限内做出客观的数据分析，并提出合理的建议。因为根本没有“过去的经验”，所以为了获得高层管理者的认可，我只能逐一确认客观事实，并用环环相扣的故事把数据统一到一起。

在客观事实当中，只有数据（数字）对任何国籍、职业经历或语言背景的人都能发挥威力。不过当然不能只把业绩数据做成图表就直

接提交上去，必须用客观的、合乎逻辑的方式证明“为什么会这样”“为什么说这样不好”，才能让对方接受“那么需要怎么做”。

例如在海外设立新公司、大规模重组、彻底改变业务流程等，这些左右公司命运的重大决策，是怎样在最高层经营会议上提出来的呢？缺乏事实和逻辑根据的方案在提交给 CEO 之前，根本就无法获得相关部门高管的批准。任何项目的方案都必须满足一个最基本的条件，那就是首先要得到相关部门高管的认可，让他们觉得“这个内容可以接受，而且也足够简单易懂，能向 CEO 解释清楚”。

如果缺少任何人都能看懂的“数据”，没有能说得通的“故事”，方案就无法在有限的时间内获得批准。如果仅凭一时灵感或者主观偏见，导致内容与事实相悖，或者提示了错误的方向，则可能造成严重的后果，包括相关的人际关系等都会变得难以收拾。尤其是在很多项目中，各部门的责任范围都非常明确，各领域之间存在着对立的利害关系。因此我需要一些技巧，保证内容必须符合事实，而且不能作为个人的意见来阐述，而是要让数字或数据来“说话”。

我并没有可以适用于任何情形的方法。不过得出结论的过程以及数据的着眼点、基本的分析方法等并不局限于特定的部门或问题。这些最基础的部分就像计算机的操作系统，能通用于很多解决问题的项目。

复杂的管理问题不一定需要高难度的统计方法。简单易懂的表达方式反而更为重要。虽然也有很少一部分项目完全不涉及数字，但大部分问题都与金额或指标等数值紧密相关。

本书介绍的思路和方法，都是数据分析、统计分析、数据科学等专业领域中最为基础的部分。在挑战公司中的一般问题时，高难度的方法未必能获得正确答案，有时反而会使我们离目标越来越远，这样的事例并不少见。

除了规模宏大的管理问题之外，日常的实际业务中也可以用到完全相同的思路、工具或流程。这些方法可以根据更多的信息及更客观

的事实，更高效地解决各种问题。充分运用数据，可以为此发挥重要作用。

如何在实际工作中使用数据

我们常说将数据应用于实际业务，但具体要怎样做呢？根据不同的阶段和水平，我们所需的能力可以分为以下 4 种。

（1）解读数据

掌握观察数据的“视点”（这是驾驭数据的基础）

（2）基本的数据分析方法

掌握通用的整理和分析数据的基本方法

（3）综合技能

将各种方法或思维方式有机结合起来，形成具有整体一贯性的解决问题的故事

（4）高级分析技能和 IT 系统

掌握运用专业的高难度方法和高难度技术进行分析的能力

有一些专门从事高难度数据分析的专家，被称为数据科学家。并不是所有的商务人士都需要这些高难度的分析技能。根据我的个人经验及我与各行业客户接触的切身体会，对一般商务人士来说，具备（1）~（3）的技能就足够了。

虽然作为数据分析的第一步，解读数据的能力极其重要，但商务人士在运用数据来解决问题时，最具有说服力和逻辑效果的是“综合技能”，即将各种数据分析结合起来，通过故事找到解决方法的能力。

本书在介绍各种基本数据分析方法的同时，将构建故事作为重点。有些章节会分别介绍单独的数据分析方法，但面对越是复杂的问题，组合运用的效果就越好。希望读者能感受到，1+1 不仅等于 2，有时 1+1 还会等于 4、等于 5。

此外，本书还会根据各章主题，结合简短的事例来介绍一些单独的分析方法和思维方式。与此同时，本书还会通过各章内容共同讲述一个"解决问题的故事"。

如果有人问我在日产公司参与的各种项目当中，有多少是可以运用本书介绍的内容来解决的，我可以断言"几乎所有都可以"。我并不是力求针对某些特定问题得出分析结果的数据科学家，也不是撰写学术论文或报告的统计学家。我是"解决问题的专家"，我的目标就是在组织当中，从更贴近实际业务和管理的立场，与相关人员达成共识，以推动工作前进并取得成果。

要实现这个目标，本书介绍的内容已经足够了。当然，在大量实践中积累经验，可以进一步提高技能。希望各位读者能身临其境地面对每个问题，与我一起思考。

那么，请大家在阅读本书时，假设自己正面临着下一页的问题。本书各章的最后会依次介绍"解决问题的故事"，提示解决这个问题的线索。

要解决的问题

近一年来，你所负责地区的汽车销售额明显下滑。你和周围的人都接受了社会上大多数人的观点，即认为其原因在于年轻人对汽车的疏远和少子高龄化的发展。

所有人的思路都停留在"这样下去可不行""整体经济形势不容乐观"的原地打转，重复从过去一直沿袭至今的挽回措施：增加拜

访客户的次数，增加邮件推送广告的数量，希望能对现状有所弥补。

终于，销售本部长感到不能再这样持续下去了，命令你提出新的改进措施。如果你的提案获得批准，将会立即得到实施。提交期限是下周周末，然而到目前为止，你还没有想好应该如何改进。

这时，你得知总部的市场调查部收集了汽车用户的市场调查数据，于是向他们要来了相关资料。果然不出所料，这一年的销售额每月都在下降。像以前一样仅靠恒心与毅力，恐怕难以应付这种困难局面了。

那么，到底应该如何解决这个问题呢？让我们现在就开始行动吧。

第 1 章
解决问题，你需要“流程”

实际工作中的大部分数据分析都只是运用分析工具来“摆弄”现有的数据。因此，虽然不知不觉中做出很多图表，但耗费大量的精力和时间，最后也不过是总结出一份报告，只是描述了“本月 XX 分店取得了优异的销售业绩”“经费的使用额在整体上有所增加”等实际情况。当然，这样很难找到有效的应对措施，或者找到造成这种现象的原因。

为了避免这种情形，我们需要了解“流程”，这是数据分析的前提。掌握了流程，才能避免见到数据就立即动手，却总是在原地打转的情况。分析的各个详细流程还会在后面的章节介绍，我们先来概观一下数据分析的整体流程（故事）。

第一个重点是“划定范围”

首先，我们要了解在实际工作中解决业务问题的整体流程，以及这一过程中涉及“数据分析”的范围。解决问题的流程从明确目的或问题开始，然后再逐步运用数据来把握现状，或者从所有数据中找到问题的关键（“A 店铺”“B 商品”“20 ~ 30 岁男性顾客”等）。

从开始运用数据把握现状到找到问题关键为止的过程叫作“数据整理”。这一阶段仅限于整理数据，还没到挖掘数据进行分析的阶段。这一阶段的特点是直接运用“总销售额”等表示整体情况的数据。

不过，所有业务都出现同样问题的情况比较少见，大多数情况是只有其中的某些关键部分才是主要问题。为此，需要逐渐划定数据的范围。

不划定对象数据，在所有数据上做文章的话，由于多个要素掺杂在一起，会很难找到问题的关键（着眼于何处、如何划定范围会在第2章介绍）。只有确定了分解数据的角度，对分解之后的要素进行比较，我们才能发现问题的关键。这样就可以从包含所有要素在内的整体数据中，划出作为分析对象的数据范围。

从锁定原因到研究对策

假设我们已经锁定了问题的关键，例如“问题是支店 A 的销售”或“问题是服务 B 的集客效果”等。那么接下来，就应该考虑“什么原因导致了这些问题”。为了找到原因所在，需要进一步缩小对象范围。关注 2 种以上数据之间的关系，有助于锁定原因。这里才是数据分析的精髓。

从上一阶段“找到问题的关键”开始，“数据整理”的工作逐渐开始变为可以叫作“数据分析”的内容。当然，只是找到原因不能算解决了问题，还要针对原因决定必须采取的措施，而且提出方案，不能只是停留在口头层面。

设定什么样的目标？数值是多少？实施需要哪些资源？只有将这些内容落实为具体的数值，决策者才会批准。方案最终获得实施，才算到达终点（实际上，之后还需要检验和反馈实施的结果）。

如果各个分析过程的内容及结果互不相干，无法形成一个连贯的故事，就没有说服力，我们需要通过整个流程，形成一个贯穿全体的故事（经过）。为此，首先要明确最重要的目的和问题，确保所有过程与操作在理论上最终都与这一目的或问题相关，这样整体的逻辑和流程才不会出现偏差。

此外，就像我在日产的所有工作一样，大多数情况下我们都必须在有限的（多是极其有限的）时间里取得某种形式的成果。如果分析漫无目的，时间转瞬即逝，结果很可能会超过期限。

把握包括分析阶段在内的整个解决问题流程，就能从全局来考虑时间的分配，了解“还有多少时间可用”“目前在整个进度中处于哪个阶段”等。

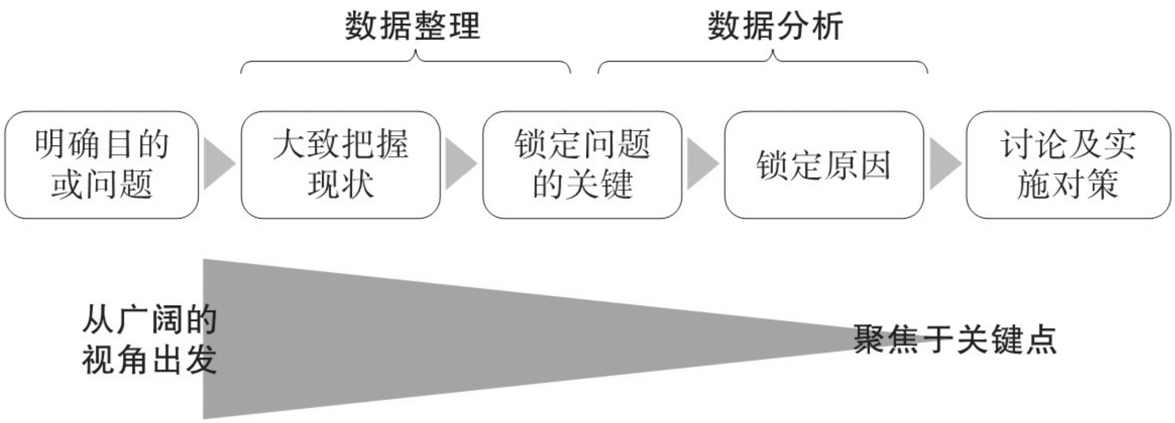
这样做的优点是可以平衡准确度与时间之间的关系，考虑为目前从事的分析分配多少时间、答案需要精确到何种程度的同时，有计划地开展工作。这样，我们到达“得出答案”这一终点的概率也会大大提高。

如果没有时间限制，我们或许可以不计成本地进行高准确度的分析，但对不是专门从事分析的普通人来说，这是不现实的。为了在有限的时间里取得一定的成果，建议大家从一开始就对整体计划和“地图”做到心中有数。

开始之前的准备工作

进入具体分析作业之前，必须先明确目的或问题，接下来就详细介绍这一部分内容（图 1-1）。

图 1-1 解决问题的流程从“明确目的”开始



不常做数据分析的人，开始可能很难迈出第一步，而是一直在原地打转。肯定很多人都有过这样的经历吧?

接到上司“数据分析”的指示，很多人会先把过去几年公司销售额的变化做成柱状图，然后再用折线图来体现每年相对上一年的增长。接下来，对不同种类的商品进行 ABC 分类。ABC 分类是指按照从大到小的顺序将销售额排序，根据不同商品在整体中所占比例的大小，将其分为“A 类”“B 类”“C 类”，这是一种排列优先顺序的分析方法。

利用手边的数据，暂且先做了这些分析。可是在这之后，人们就会陷入停滞不前的状态，“接下来该怎么做呢……”

会出现这种情况，是因为分析者只关注了代表结果的数据，而没有具体考虑分析的目的或问题。

在解决问题的过程中，数据分析的目的是针对“原因是什么，需要采取哪些行动”等问题得出结论。销售额、利润等数据体现了公司业绩的最终结果，这类数据容易收集，也更容易得到人们的关注。但是仅盯着结果数据进行分析，无法实现“改进结果”“解决问题”等目的。

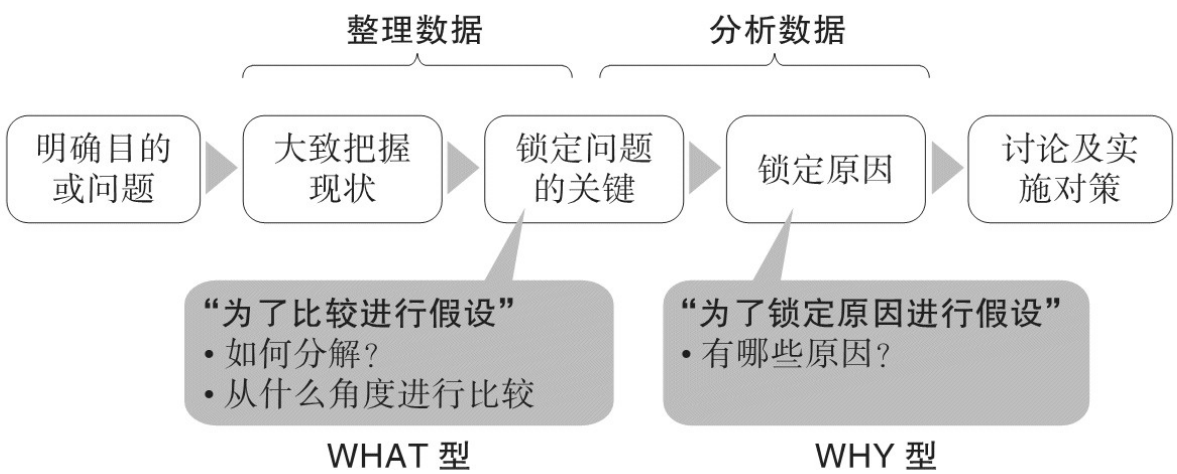
抱着“先从容易收集的数据开始统计”的想法进行分析，这本身就是错的。正常情况下，首先应该确定分析的目的。但是忙于日常业务的商务人士在面对数据时，却往往会想要“先做成图表看看”。他们明知没有意义，却还是不由自主地这样做。

这样一来，无论怎样分析，大多数工作都是徒劳，效率非常低。在我举办的数据分析培训中，也有很多学员无法正确地迈出数据分析的第一步，不断重复错误。不要从分析（方法）开始，而要从明确目的或问题着手，这一点是最重要的。

数据分析需要“假设”

确定目的或问题后，才能决定假设、方法、所需数据等具体内容。进行数据分析，“假设”可以发挥重要作用。如图 1-2 所示，在解决问题的过程中，有两个环节需要假设。

图 1-2　分析流程中必不可少的“假设”



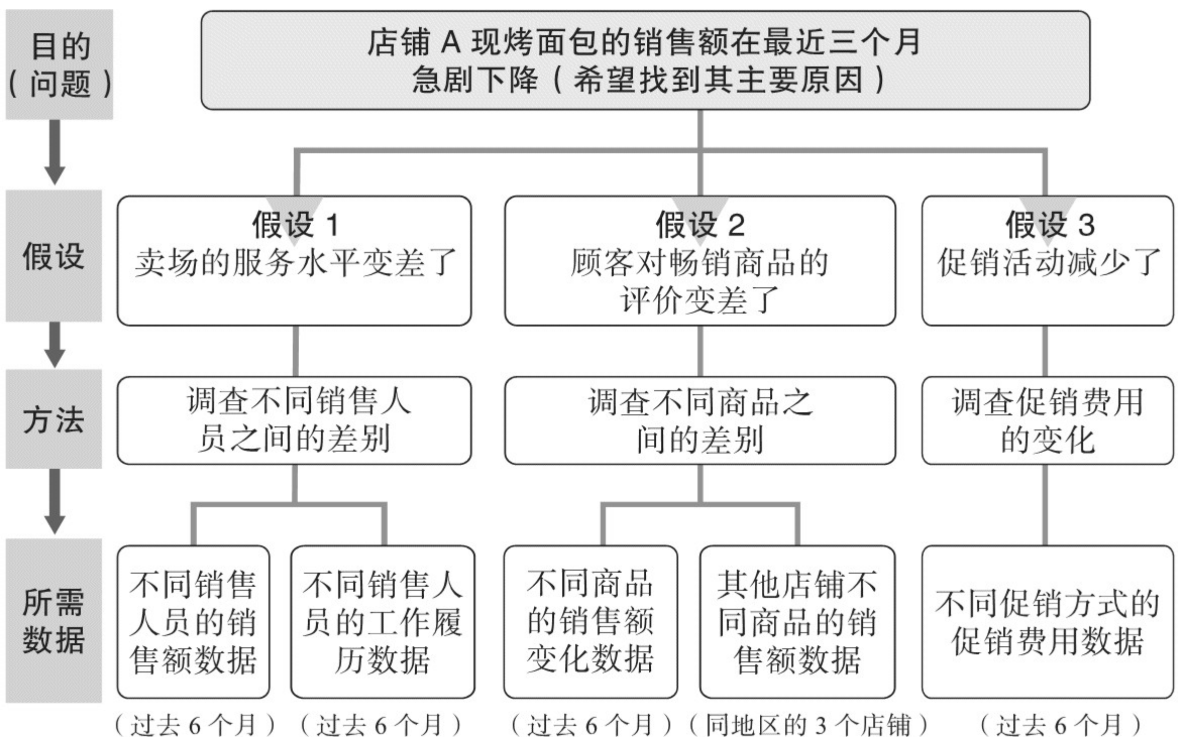
首先，在“锁定问题的关键”时，需要通过“WHAT 型假设”对数据进行分解和比较。此外在“锁定原因”时，还需要通过“WHY 型假设”来列举出候补原因。

关于这两种不同类型的假设，第 2 章将会介绍具体事例，在此仅用一个例子来说明如何通过假设得知“分析什么，如何分析”。假设我

们使用 WHY 型假设，来探讨店铺 A 的现烤面包销售额在最近 3 个月急剧下降的原因。

为了整理思路，可以制作一个如图 1-3 所示的图表。按照逻辑关系，将目的或问题、假设、方法以及所需数据从上至下依次连接起来。

图 1-3　通过假设得出所需的方法及数据



假设听起来好像很难，也许会让人有些抵触。其实我们也可以把假设替换成“着眼点”来考虑。

在这个事例中，可以考虑以下三种假设（为了便于参考，在此做了简化）。

假设 1：卖场的服务水平变差了（服务的原因）

假设 2：顾客对畅销商品的评价变差了（商品的原因）

假设 3：促销活动减少了（促销的原因）

大家可以发现，这里并没有列举什么特别高深的内容。重要的是，要将这几点用语言表达出来，确保任何人都能看懂。括号内的原因为着眼的关键词。

如果觉得最初的假设挖掘得还不够深入，也可以进一步反复思考“为什么”，继续深入挖掘第二层、第三层假设，从而找到更为具体的问题（原因）。这里不再详细介绍深入挖掘的方法，但请注意：并非所有的情况都是只有一层假设就够了。

接下来，要确定检验各假设的方法及所需数据。如果检验假设 1 的方法是“调查不同销售人员之间的差别”，那么所需数据就是“不同销售人员的销售额数据”和“不同销售人员的工作履历数据”。对假设 2 也一样，如果将方法定为“调查不同商品之间的差别”，则需要“不同商品的销售额变化数据”和“其他店铺不同商品的销售额数据”。按照这个步骤进行准备，就可以大大减少面对数据不知所措的场面。

做完了这些工作，接下来总算到了分析数据的环节。我们要根据数据来检验假设。要确定对问题来说，哪种假设是正确的、哪种假设的影响最大。做到这一步，就能够发现问题出在哪里、应该采取哪些对策，或者是否需要继续深入调查或分析。也就是说，这样就可以采取下一步措施。

研究假设的效果并不仅限于此。它还有助于制作提交给公司内部或外部的演示资料。在开始分析之前，将“问题”“假设”“分析结果”“结论”等必须以符合逻辑的方式展示给其他人的主要内容或流程确定下来，这样可以提高工作的效率。

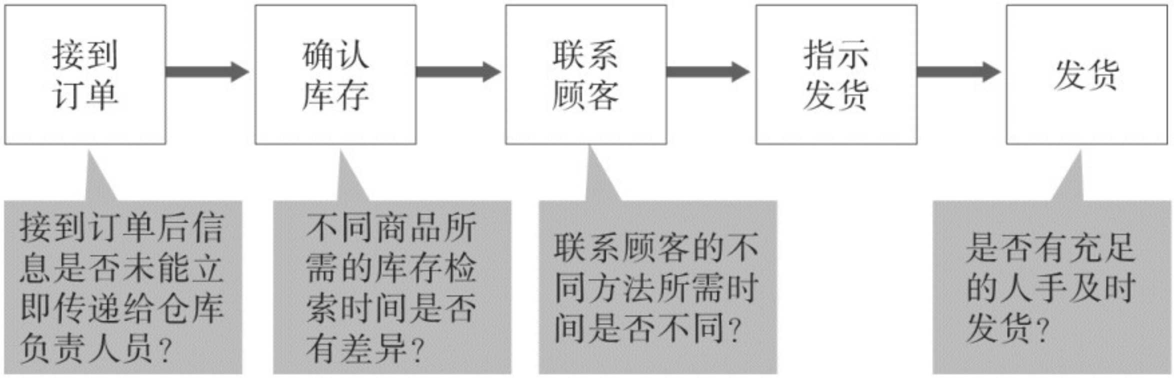
接下来再介绍一下建立假设的方法。人们建立假设时，很容易依赖工作中的常识、自身的经验或直觉，可能会无意识地排除一些重要

的假设。因此很多人会担心自己的假设是否存在偏差、是否视野过于狭窄。

我们既然采用了数据分析这种科学的研究方法，也就应该尽量客观地提出假设。为此，可以请别人对自己的假设进行确认，也有一些可以用来自己检查的方法。方法之一是运用“框架”作为思考问题的线索。这里的框架是指商务人士都经常会接触到的流程图、损益表及 4P 营销理论等（图 1-4）。

图 1-4 可用于假设的“框架”

流程图示例（目的：缩短从接到订单到发货的时间）



损益表（P/L）示例（目的：防止利润继续减少）

项目	假设
销售收入	单价或数量是否减少了？
销售成本	人工费、原料费是否提高了？
销售总利润	
销售费用及一般管理费用	促销费用是否增加了？
营业利润	
营业外收入 / 支出	最近的营业外支出是否有变化？
净利润	

4P 营销理论示例（目的：防止销售业绩继续下滑）



这些方法可以帮助我们整理思路。针对“缩短从接到订单到发货的时间”的目的，可以通过流程图，整理出“接到订单” “确认库存”“联系顾客”“指示发货”“发货”等各工序中可能存在哪些问题。

比如，在接到订单的阶段，可以假设“接到订单后信息是否未能立即传递给仓库负责人员”；在确认库存阶段，可以假设“不同商品所需的库存检索时间是否有差异”等。这样可以更容易针对具体对象提出假设，也便于确定采用的方法和所需的数据。

损益表也有助于建立假设。面对“防止利润继续减少”的问题，可以针对“销售收入”“销售成本”“销售费用及一般管理费用”等项目，分别提出假设。例如，针对销售收入可以假设“单价或数量是否减少了”，针对销售成本可以假设“人工费、原料费是否提高了”等。

此外，我还建议运用 4P 营销理论的思维方式，将问题按“产品”“渠道”“促销”“价格”进行分类。如果问题是“防止销售业绩继续下滑”，则可以从这 4 个角度来提出假设。这样就比较容易整理出如“产品竞争力下降了吗”“因断货而错失销售良机了吗”“促销的效果减小了吗”等问题。

用框架来查缺补漏

前面介绍了框架的作用，但这并不说假设必须完全符合框架。使用框架，是为了检查假设是否有遗漏。此外，框架还有助于明确“可视的范围”和“不可视范围”。它可以帮助我们认识到，“自己目前正在哪个范围进行研究”。这一点极为重要，因为它直接规定了数据分析的范围。

例如对产品非常熟悉的人，就有可能将注意力只集中在产品上，而不太会考虑渠道或价格。即使是十分了解业务流程的人，对于销售收入、销售成本等财务方面的假设，有时也会出现遗漏。因此，利用

框架提炼出需要考虑的关键词，再次确认自己的视野是否受到局限才是万全之策。

此外，遗漏并不都出于无意识的疏忽。我经历过的公司内部改革中，对一些需要伴随痛苦的改革、改进或问题，相关人员有时不一定会爽快地提供、明示所有的知识或信息。尤其是对自己不利的信息，有人甚至会刻意隐瞒。

这种情况下，就必须依靠其他通用信息（框架等），实现尽可能完善的分析。尤其是对方远在国外的情形下，难度就会更大。因此我们需要具备克服这些困难的能力和技巧。

接下来再介绍一下提出假设时的其他关键。从能影响各项假设的驱动视角来思考问题，更便于找到下一步应该采取的措施。驱动是指“自己能够控制的数据”。因为根据分析数据的结果，找到解决问题的措施，才是大多数问题的最终目标。

根据应该采取的措施来提出假设的做法是本末倒置，但如果提前意识到之后应该采取的措施，则可以避免出现最糟糕的情形，即“分析后却没有找到应该采取的具体措施”。能否做到这一点是重要的分水岭，决定了我们的工作只是以分析告终，还是能够解决问题。

看到这里，也许有的读者会觉得，“假设是很好，但是在数据分析之前要做的工作太多了。”这一点其实不用担心。因为我也并不是每次进行数据分析之前，都要制作类似图 1-3、图 1-4 的图表。

在尚未习惯数据分析的阶段，制作图表是一种行之有效的方法。但实际上很多时候，这个工作只要在头脑中做就可以了。制作图表可以提高数据分析的效率，更容易实现目标。但如果成为负担，害得我们为此而顾不上关键的数据分析和解决对策，图表就失去了意义。

此外，即使提出了思路清晰的假设，确定了所需的数据，在实际工作中却很少会遇到所需数据全都摆在眼前的情况。不过即便这样，我们也不应该根据“现有数据”来开始所有的工作。只在“现有数据”的

范围内进行分析或者解决问题，会妨碍我们发现问题的本质及有效的对策，很可能导致我们的工作仅以“数据游戏”告终。

没有数据就要不计代价地花费时间或金钱去收集数据，这种观点没有错，但在实际工作中却是不现实的。因此商务人士需要设法借用相同或相近的数据来代替理论上所需的数据。

例如，假设我们需要“工作繁忙程度”数据。如果对所有员工就最近一个月的繁忙程度进行问卷调查，应该能够获得包括主观评价在内的数据。但是如果不具备进行问卷调查所需的时间或人力，就应该寻找与“繁忙程度”近似的数据。这样就可以拓宽思路，用“每月人均处理票据的数量”“平均加班时间”“日平均工作时间”等常规的管理数据进行分析。

我在日产负责中东及近东市场时，曾经需要预测未来的市场需求。当时没有日产汽车的需求预测数据，我必须自己收集可能与之相关的数据，并用数字来证明自己的预测。原油价格、外汇汇率、对象市场的 GDP（国民生产总值）等数据可以反映出中东及近东市场的特征，并且在相关机构的官方网站上都能够查到。于是我将这些数据与日产过去的业绩进行对照，从中选出可以用上的内容，尽量从自己能够获得的数据中挖掘出更多的信息，反复摸索试错。

这种努力已经不是分析方法的问题，而是依靠个人创意和毅力获得成果的尝试。这次经历也使我改变了“没有合适的数据就无法分析”的观点。这件事坚定了我的信心，对任何看似有用的数据，都要想方设法从中挖掘信息，分析出尽可能正确的结论。

开端决定了结论的质量

下面针对最初的“设定问题”做一些详细介绍。从顺序上来看，“设定问题”位于“假设”之前，直接决定着分析的质量。对所有问题来说，

最初的“问题定义”都决定了后面工作的成败。

针对实际业务的数据分析也是同样的道理。例如，假设下属想按照下面的步骤进行分析，您会在哪些地方指出他们的漏洞呢?

问题：过度降价导致洗发水 A 的年度利润下降

↓

假设 1：某些店铺过度降价

假设 2：集中在特定时期过度降价

假设 3：与竞争对手展开价格战导致过度降价

↓

假设 1 所需数据：每个店铺的实际降价数据

假设 2 所需数据：每月实际降价数据

假设 3 所需数据：竞争对象产品的价格变化数据

收集数据进行分析的结果显示，不同店铺之间确实存在降价幅度的差异，不同月份也可以看出某种程度的季节性变化，此外竞争对手也在挑战降价极限。检验的结果，就是所有假设都不是毫无根据的。

不过，大家是否看到这个看似顺理成章的结果中潜藏的风险了呢？其最大的风险是设定问题时，分析者将思考范围狭隘地限定在“自己设想的范围内”。尤其是当分析者拥有具体的直觉和经验，并对其深信不疑时，这种倾向就表现得更为显著。很多时候，分析者并没有意识到，他们进行分析仅仅是为了验证自己的想法。

如果分析的结果能够支持（或接近）自己的想法，分析者就会大大满足，对风险就会变得极为迟钝。他们往往会产生一种错觉，认为分析取得了成功。这样一来，耗费时间与精力所做的分析就都成了徒劳。

如果最初就对自己的想法坚信不疑，即使只能从分析结果中获得零散的支持，人们有时也会用“自己的故事”去填补其欠缺之处。当我们遇到看似完美的问题设定、完美的假设、完美的分析结果时，就应该怀疑它是不是属于这种情况。这种分析大多是一开始就有了结论，未必具有很大价值。

那么，这个事例中具体存在哪些问题呢？答案是，在问题设定中，就已经包括了“原因”（在这个事例中为降价）。越是具有丰富经验及敏锐直觉的人，越有可能在设定问题时就已经想好了针对原因需要采取的“对策”。

这项实际业务技能本身很了不起，但客观分析的一个重要步骤，就是要有意识地回避“思考的近路”。因为在设定问题时，分析者的思考还无法超出“降价”的范围。

从刚才的事例来看，所有的工作都变成了收集与降价有关的数据，从而得到分析的结果。要避免这种情况，在定义问题时，应该尽量具体地思考“怎样才是最理想的状态”。

这样一来，就可以通过确认希望实现的理想状态，来扩展自己的视角，能避免思路立即集中到具体原因上。

我过去就曾遇到过一位“在这一行业潜心钻研 30 年”的部长级人物，由于他断言“这绝对是因为 XX”，使我接下来费尽周章，才得以进一步探究更接近本质的问题。在这种情况下，最重要的是不要轻易放弃，要用客观数据解决问题。很多时候，都能找到以前未被觉察的问题或原因。

在刚才的事例中，对“怎样才是最理想的状态”来说，答案是“最理想的状态是洗发水 A 的年度利润保持一定水平之上”。虽然也可以采用相反的方式，从一开始就定义“问题在于 XX”，但在定义“XX”时，人们往往容易加进自己设想的原因或对策。

为了避免这种情形，建议首先从正面确认理想的状态是什么。这样可以有意识地将“本质上的问题”与“原因”暂且分开考虑。

例如，如果将问题定义为“成本过高”，那么思路就会被限定在“降低成本”的范围之内。但如果考虑“希望达到怎样的状态”,则较易形成“不要挤压利润”的想法。这样才能确定“增加收入”的最终目标，与只考虑成本相比，视角就会更加宽阔。

在此基础上定义问题，即现状尚不满足这个理想状态。在该事例中，问题是“洗发水 A 无法保持一定水平以上的年度利润”。可见问题原本就是指现状与理想状态之间的差距。因此不知道理想的状态是什么，就无法确定问题。跳过这一步展开分析，会带来巨大的风险（徒劳无功）。

当然，我们使用数据解决定量问题，就尤其需要考虑如何用数字来体现现状与理想状态之间的差距。用数字定义问题，具有以下两个优点：（1）可以明确用哪项指标来衡量；（2）可以定量地与其他人一起确认问题的严重程度及最终目标。

接下来，要思考为什么现在没有达到理想状态。这是（WHY 型）假设之一。这样思考，就不太会认为原因“只是降价”或“只是成本”了（图 1-5）。

图 1-5 根据理想的状态及其与现状之间的差距找到问题



作为示例，我们列举了除“降价”以外的其他原因（假设）。由于问题是更高一层、更接近本质的“利润”方面的内容，那么与利润相关的所有要素，都可以成为假设。

顺便提一下，如果不是为了解决问题，而是从最开始就有希望实现的明确目标的话，情况就完全不同了。例如在“为了将产品投放到某个市场，想用数据证明市场规模足够大”的情况下，就可以直接将其设定为目标，提出能够检验这个结论的假设。

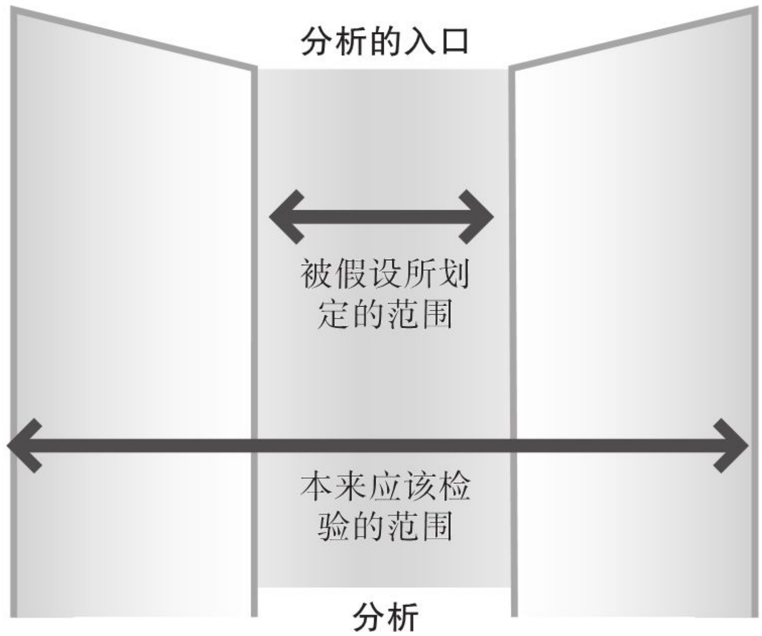
开始以后很难再扩大视野

开始数据收集或分析之后，有时也会想“还是再看看别的数据”。不过这种场合一般只是凭着偶然的念头来扩大范围。开始分析之后，要想重新扩大视野，改变分析范围，无论在时间上还是在心理上，都具有很大难度。

因此，可以毫不夸张地说，能否在分析之初从逻辑上设定合理的范围，直接决定了之后的分析质量。研究假设的风险之一，是提出假设会缩小思考及分析的范围，可能导致即使得出一些分析结果，却都不是最佳的答案（有时还会导致无法得到想要的结果）。从这个意义而言，决定胜负的第一扇门在于假设。

邂逅预料之外的分析结果，才能帮助我们找到更接近本质的答案。只在最初预想的范围内进行分析，得出意料之中的结果，就无法充分发挥数据分析的作用。大概很少有人意识到，在分析流程的入口处还会有这样的陷阱吧（图 1-6）。

图 1-6 设定更宽的检验范围（思考范围）



思考问题之“外”的问题

在日产工作时，对高层领导或上司布置的问题，我从不就事论事地只看问题本身，而会留意“问题之外的问题”，试着从高出自己一两个级别的水平进行思考。很多时候我向提出者进行确认后，又重新定义了问题。每次遇到这种情况，我都会感到自己的视野过于狭窄。不过这样的经历多了，就逐渐培养出在接手项目时，从开始阶段就扩展思路的能力。

现实工作中，即使明白这个道理，仍有很多人会觉得，与其自己刻意扩大问题范围，逐一梳理原因，还不如只选取自己和周围人都能认同的问题，用不疼不痒的结论敷衍了事，这样工作起来才轻松。

尤其是在既要负责日常的本职工作，又需要解决问题的情况下，“差不多就下个结论，尽快处理完”的心理就会占上风。从零出发

精确分析的理想论与现实中的时间和精力相互较量，二者之间需要达到平衡。当然，这个平衡点的位置会因人而异。

我也曾经历过很多类似情况，不过后来我认识到，从平时就养成高效处理日常业务的习惯，确保留出一定时间的技能和意识也是解决问题的能力之一。大多数情况下，对比我们高出一两个级别的人来说，分析者采用了怎样的研究方法，他们只要看看结果就会一清二楚。

此外，如果一个人总是在差不多时放弃努力，那他将永远无法掌握具有更高价值的工作技能。事实上，只是如何设定问题这一项，就完全能够体现出分析者的工作态度。

看到这里，相信您已经发现了，高质量、高效率的数据分析并不只取决于“数据分析”本身的质量。也就是说，在实际动手处理数据之前，应该先缜密地思考，这一阶段的内容和质量基本上决定了后面分析的质量及整体故事的有效性。

从接受提案的一方来看，他们认同的并不是数据分析的内容和方法，而是以清晰明确的、能够说服别人的问题或逻辑前提为基础提出的建议。解决问题的目标是准确地锁定主要原因，让相关人员接受自己的建议，在获得认同的基础上“推动其采取措施”。否则，无论展示出多么高超的分析方法，也无法实现这个目标。

「解决问题的故事 1」

问题：汽车销量下滑，怎么办?

第一步：明确计划与现实之间的差距

“汽车销量下滑”，这是公司内部公认的事实。那么是不是就可以把这句话设定为问题，然后直接进入数据收集和分析阶段呢?

如果这样立即开始工作的话，恐怕会由于没有与相关人员确认“下滑”的准确含义、是否可以直接作为问题、对公司事业具有多大影响等信息，导致分析者依照自己的主观感觉来决定优先顺序和判断标准。

定量评价或判断数据分析结果时，尤其需要客观的判断标准。如果标准含糊不清，我们花费大量时间和精力得出的分析结果就有可能在最后关头被全盘推翻。

这里首先需要在定量把握问题的同时,把握大致的现状。因此我们必须对问题及其主要指标做出如下定义，并与相关人员进行确认。

“汽车销售额自一年前开始逐渐减少，与上一年度的月平均销售额相比，本年度最近一个月降低了大约 15%。这意味着与本年度计划相比，全年预计将有约 10% 的目标无法实现。”

也就是说，首先需要确定，用“销售额”作为衡量汽车销量减少程度的指标。

接下来，为了确认这就是“问题”，还需要明确现状与计划之间的差距。即从最开始就要明确没有达到计划的“现实”与“理想（计划）”之间的差距。这样一来，接下来就可以讨论要如何做、做什么、需要改进多少的问题了。

做了这些工作之后，我们才算站到了起跑线上，可以为解决问题而构建“故事”了。除此以外，我们还明确了前进的方向。这样，接下来就不会偏离目标，可以有效地进行分析了。

第 2 章
分解数据，找到“问题的关键”

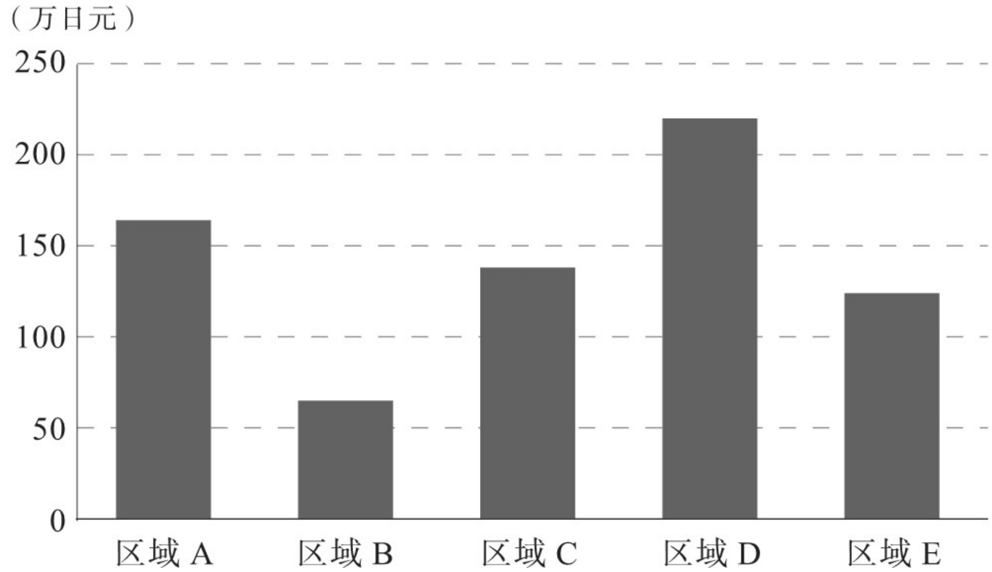
明确目的或问题、大致把握现状之后，就进入提出假设、确定问题关键的阶段。那么，具体应该如何找到“问题的关键”呢?

“问题的关键”是指从诸如“东京支店”“产品 B”“20 ~ 30 岁女性顾客”等切入点找到的引发问题的关键部分。在这个过程中，把握数据和分解数据的视点很重要。因为要锁定“问题的关键”，需要准确把握数据所显示的信息，在此基础上对分解的结果进行比较，从而锁定关键点。

例如，面对“最近销售业绩下滑”的问题，如果不考虑“用哪些数据、看哪里、怎样看”就开始分析，就可能花费了大量时间和精力，却仍然无法获得理想的结果。

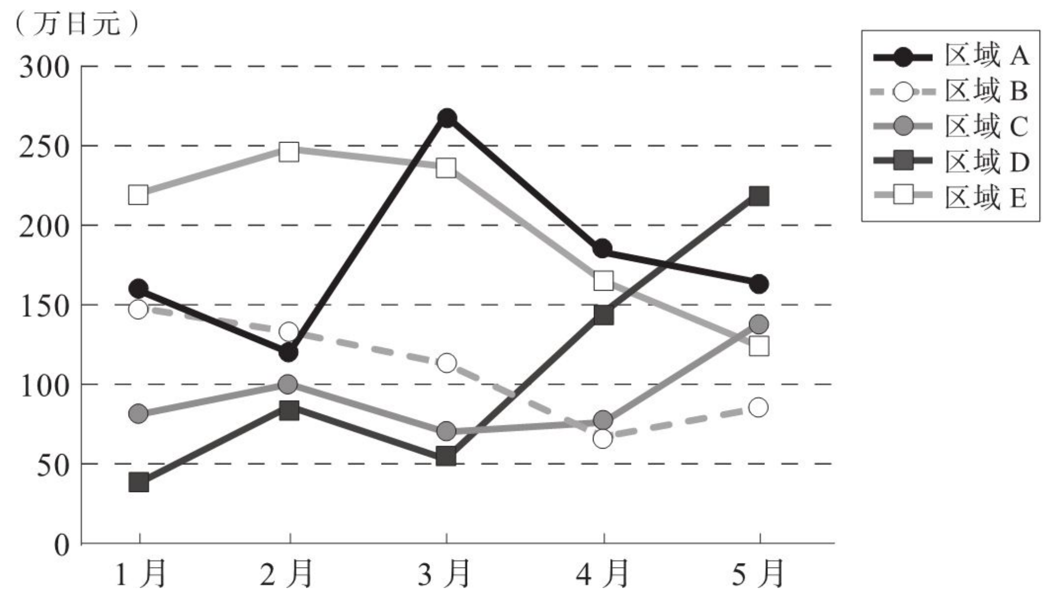
我们来看下面的例子。图 2-1 是根据手头的上个月销售数据，假设“某个区域存在问题”，用不同区域的日平均销售额做成的柱状图。如果有人根据这张随处可见的图得出结论，认为“区域 B 存在问题”，他的上司一定会提出质疑。

图 2-1 不同区域的日平均销售额（某个月）



“这只是上月的数据。其他月份是什么情况？今后预计会变成怎样？”对此，可能很多人都会（在心里）大喊：“早说啊”“那要怎么办”。然后不得不再去查看过去几个月的销售额，重新进行分析（图 2-2）。

图 2-2 不同区域的日平均销售额（过去 5 个月）



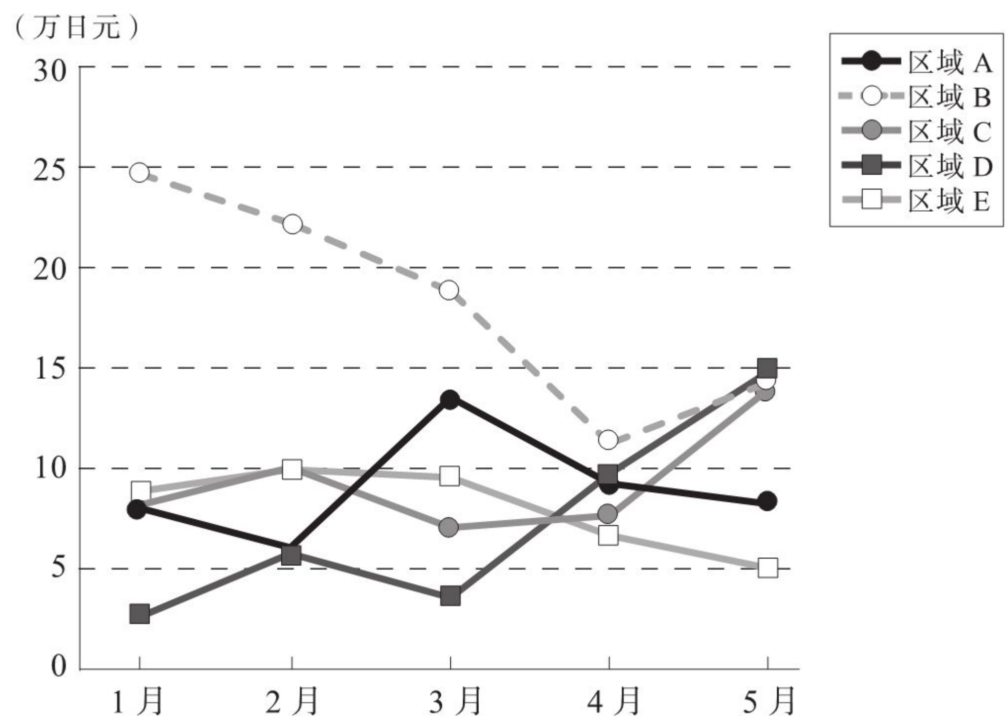
从这个折线图反映的趋势来看，似乎可以认为“区域 B 的业绩的确不好，不过区域 E 的减少趋势也需要注意”。把这个图表拿给上司，结果又被追问：“你比较的是日平均销售额，但这些区域能直接比较吗？”如果每个区域的店铺数量不同，单纯比较销售额就没有任何意义（图 2-3）。

图 2-3 每个区域的店铺数量

	区域 A	区域 B	区域 C	区域 D	区域 E
店铺数量	20	6	10	15	25

用每个区域的销售额除以店铺数量，重新把各区域的单个店铺平均日销售业绩制成图 2-4 所示的图表。从这个图可以发现，与其他区域相比，区域 B 的销售业绩并不差。至此，已经重新分析两次了。

图 2-4 每个区域的日平均销售业绩（单个店铺）



其实这些工作都只是“数据整理”，并没有通过“数据分析”来寻找导致销售额减少的原因及其对策。也就是说，做到这里才算是通过分析数据锁定了应该深入挖掘的对象（问题的关键），站在了分析的起跑线上。

可能很多人都曾经有过类似的徒劳经历，为什么会发生这种情况呢？其原因在于，分析者没有确立把握数据特征的视点，就开始着手整理数据。因此，每次只对被质疑的问题继续挖掘，然后又会被从其他视点追问，如此反复。

每当上司或者其他人指出一些问题，分析者都会觉得“他们说的确实有道理”。然而事实上，如果最初就从全面的视点把握数据，这些无用功和错误都可以避免。如果不在数据分析的“地图”上规划好到达终点的路径，而是先走起来再说，就可能会走弯路，甚至无法到达目的地。

“趋势”视点和“快照”视点

前文的例子中缺乏“把握数据（捕捉数据特征）的视点”的问题，在很多情况下可以通过“趋势”和“快照”两个视点来解决。缺少其中的任何一个，论点都容易出现遗漏，导致例子中被上司质疑的情形。因此，我在做分析时也一定会确认是否涵盖了各种视点。如果判断分析中不需要某个视点，我则会确认其原因。

“趋势” 视点可以捕捉一段时间内的变化，也被称作时间序列。通过数据观察变化经过，可以得知有哪些流程，以及形成目前状况的经过。关注过去的变化，预测将来的动向，可以带来飞跃性的效果提升。因为从时间的变化过程中，有可能找到问题的关键或原因所在。重点挖掘这些内容，找到所需信息的概率就会增加。

“快照” 视点是截取某个期间的情况。用指标（平均值等）体现该期间的大小、比例和分布情况等，可以轻松地把握大致情况，或进行比

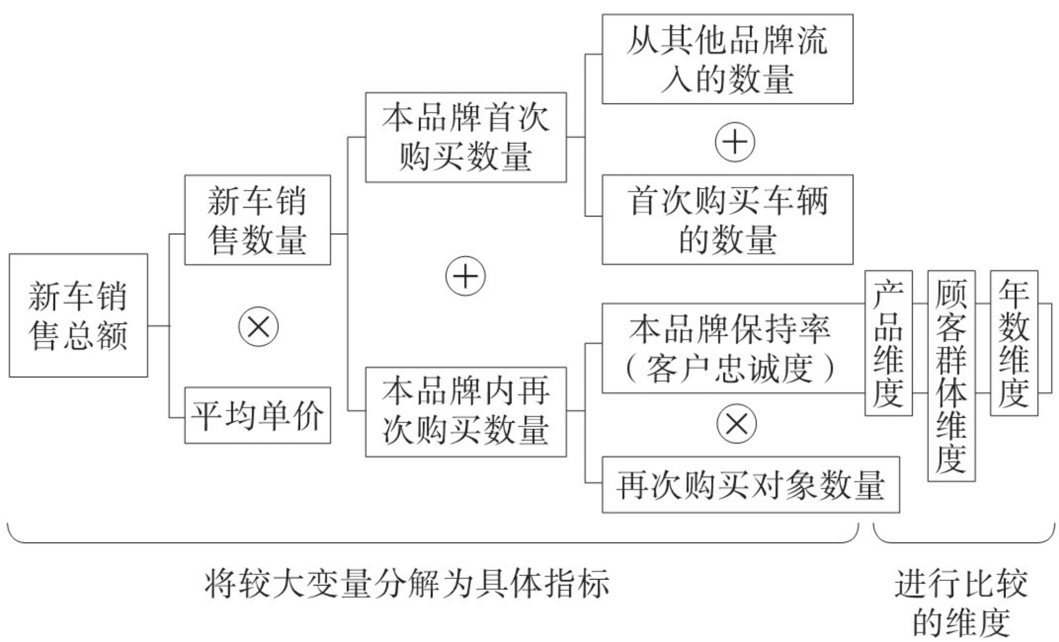
较，从而确定该因素对问题的影响程度。例如，区域 A 的销售额的确出现了大幅度下降（趋势），不过如果通过“快照”发现其年销售额规模不及整体的 1%，我们就会毫不犹豫地将其从深入挖掘的对象中剔除出去。

运用“趋势”和“快照”的视点把握了数据的整体情况之后，还可以运用相同的视点找到问题的关键。要做到这一点，可以根据相对定位进行评价，例如“销售额高于其他区域”“不良品多于 B 产品”等，这是一条捷径，即通过比较发现数据与其他因素之间的相对差异。

“WHAT 型假设”

不过，对较大范围（包括多个要素）的数据，即使进行比较也很难找到差异。此时需要运用“四则运算”来分解数据，使其变得更为详细和具体。这就叫作“WHAT 型假设”（图 2-5）。

图 2-5 建立 WHAT 型假设



在此基础上，接下来需要关于“维度”的思路。分解数据之后，会得到某个指标或者变量（例如：客户忠诚度）。我们需要考虑从何种维度对该指标进行比较，才能找到问题的关键。在依赖数据之前先进行推测，也可以说是 WHAT 型假设的一部分。

这样一来，只要熟悉业务内容，人们就会想到“区域”“产品”“顾客年龄”等许多维度。从中优先挑选与问题联系最紧密，最能体现差异的维度进行比较，便可以找到问题的关键。例如从产品维度对保持率进行比较，可以发现产品 X 明显高于其他产品，这就是问题的关键。那么，想要找到问题的关键，需要对哪些指标进行比较呢？

很多事例按大小或比例对销售额、揽客人数等加以比较，此时必须注意原始数据的波动。因为，尤其是用平均值代表整体数据的大小时，数据波动的信息就会完全消失。

找到“问题的关键”所需的视点可以用图 2-6 的矩阵图来表示。这个矩阵图可以防止比较时遗漏重点问题。我们并不一定要机械地查看矩阵图内的所有项目。不过大家需要知道，不事先考虑好“为什么要计算这个指标”“希望由此了解什么”，而是漫无目的地“姑且”考察数据波动，就无法让数据分析有的放矢。

图 2-6 利用矩阵图整理比较的重点

	“快照”视点	“趋势”视点
大小的维度		
波动的维度		

接下来，分别对“大小或比例”和“波动”做详细介绍。

表示大小的“平均值”

应用最广、最具代表性、最便于计算的统计指标大概要算平均值吧。可能有人觉得“不就是平均值吗”，不过可千万别小瞧了它。平均值虽然用起来很方便，但却有一个陷阱，让很多人都陷入其中。

为什么平均值可以广泛用于各种场合呢？我们首先从积极的方面来考虑。单纯地罗列数据的话，人们很难从中获得信息。数据越多，就越会超出人们能够处理的极限。

此时，我们可以制成图表，或者取整体的平均值。这是为了将整体归纳为图表或者平均值等“一个对象”，便于人们理解和认知。尤其是平均值能够作为体现数据“大小”的代表值或标准值发挥作用。因此，平均值这个指标用途广泛、非常便利。

例如，让同一个小组对产品 A 与产品 B 进行评价。与其列出每个成员的评价结果，不如计算出能够代表整体水平的平均值。例如得知产品 A 的平均分为 55 分，产品 B 是 87 分，人们就可以把握每种产品的大致情况，将产品 A 与产品 B 进行比较也会更为简单。此外，如果不想关注销售额每天的细微变化，而是希望了解销售额每个月的大致变化，使用平均值会更方便。

因为对于想了解每月变化的人来说，每天的数据或每小时的数据只不过是过剩信息。这种情况下使用平均值，便能剔除那些过剩信息。

像这样，将大量数据归纳为一个平均值，能为实际工作带来很多方便。而且无须额外说明，谁都理解其含义，平均值可以说是一个万能的工具。

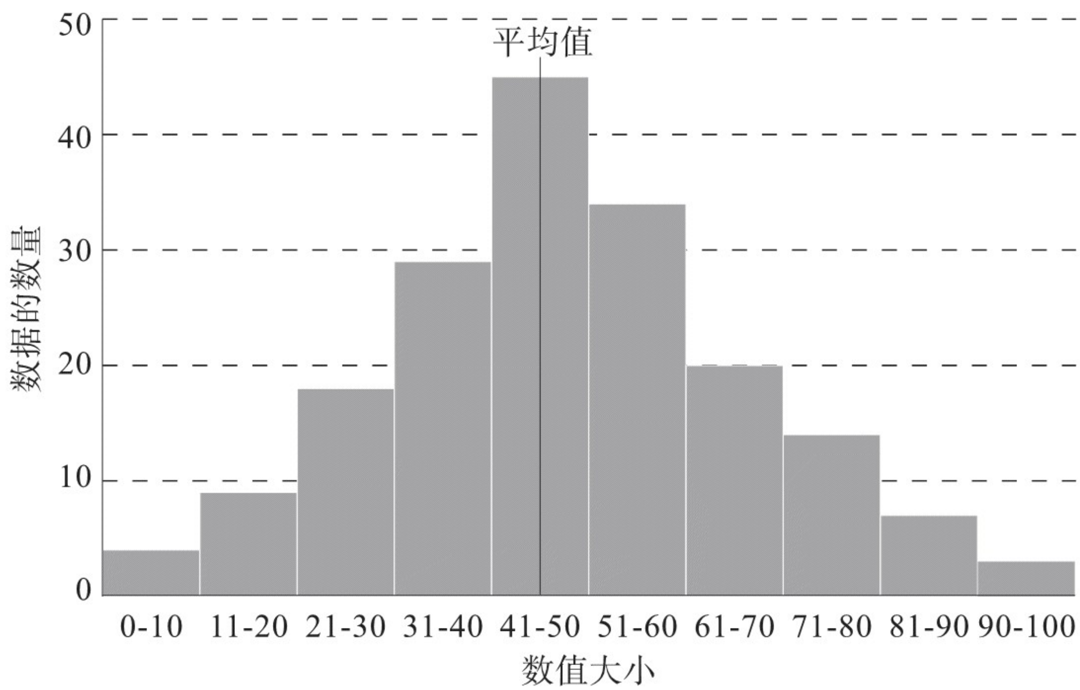
不过虽然谁都用过平均值，但面对“平均值究竟是什么”的问题，却有很多人不知如何作答。

“好用”背后的陷阱

大多数人可能会回答平均值是处于所有数据正中间的数值，或者是代表值、标准值。但很少有人会考虑“平均值真的能够‘代表’数据，或者真是‘标准’的数值吗？”有人认为平均值代表了数据整体的大小，那么他们的头脑中想象的一定是图 2-7 所示的情况。

图 2-7 是直方图，能够显示位于各个区间（横轴）的数据分别有多少个（纵轴），体现了数据的分布方式。在这个示例中，平均值处于所有数据的正中，与平均值相等或接近的数据最多。其他数据的个数随着对平均值的偏离而逐渐减少。这种情况下，平均值确实可以被称作代表值或标准值。

图 2-7 直方图示例



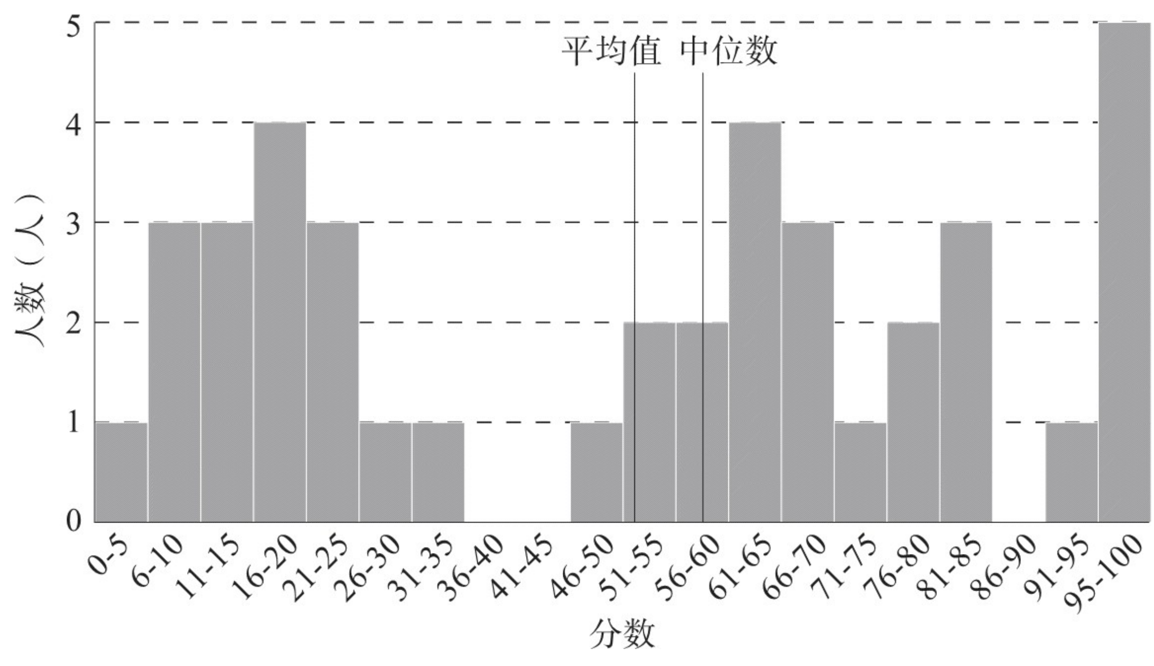
但是，如果毫不怀疑地把类似图 2-7 的情况作为前提来使用平均值，就有可能陷入意想不到的陷阱。因为未必所有数据都会呈现出同样的分布方式。

平均值不能代表总体

回忆学生时代，得知自己的考试成绩高于班级平均分时，大家是否会觉得“排在中等以上，还不错”？不过，如果您对这句话没有感到任何疑问，就需要注意了。此外，如果听到某国的平均寿命是多少岁，就认为这个国家的人都是在平均寿命前后去世，这样的人也同样需要注意。

图 2-8 是一个 40 人班级的考试成绩分布情况。其平均分为 51 分，但位于所有成绩正中间（即第 20 名与第 21 名之间）的却是 57.5 分。也就是说，得了 55 分的人虽然高于平均分，却并未排进班级的“前一半”。请注意，图 2-7 以平均值为中心，数据个数逐渐减少，而图 2-8 的数据却呈现出截然不同的分布方式。

图 2-8 考试成绩的分布



在图 2-7 的情况（往往是很多人所想象的情形）下，可以说平均值就是代表值。但在图 2-8 中，平均值周围并没有集中很多数据。因此，

按照数值大小顺序，位于中间的数值（中位数）也偏离了平均值。这种情况下，就不能把平均值看作代表数据的值。

之所以产生这种现象，是因为平均值会受到离群值的影响。而且，数据波动越大，离群值就越多。其结果就是，“平均值未必是代表值”的可能性更大。例如平均寿命也是一样，假设某个国家 5 岁以下的婴幼儿死亡率较高，那么除非有很多高于平均寿命的长寿者，否则计算出的平均寿命就会偏低。寿命集中分布在低年龄段和高年龄段两个区域，平均寿命处于其中的某个位置。这种情形下，在平均寿命前后去世的人数就相对较少。

再看一个比较极端的例子。假设对 100 个人进行问卷调查，其结果为有 50 人选择“1.极不赞同”的,而其余 50 人则全部选择“5.极为赞同”（图 2-9）。

图 2-9 对问卷调查的回答



这种情况下的平均值是多少呢？平均值应该是 3。然而却没有一个受访者选择了“3. 既不赞同也不反对”。那么这里的平均值就没有任何意义，反而还可能会掩盖人们的意见集中在 1 和 5 的事实。平均值的便利之处在于可以不用处理全部数据，只关注由整体得出的一个数值即可，但其缺点是牺牲了原始数据。这是一个很大的陷阱。

我在日产工作时养成了一个习惯，对只用到平均值的提案抱有怀疑的态度。有些人只根据平均值便断定“数值大，很好”“数值太小，不行”，这样的评价过于肤浅，其背后可能隐藏着更多信息。假如没有认识到这一点，我一定会错过很多重要的信息。

那么，有没有一种方法，可以大致判断能否将平均值视为代表值呢？根据直方图的视觉效果，可以确认到数据的分布方式和平均值的位置。但在实际工作中，不可能每次都用这个方法来确认。

了解“中位数”

这种情况下，还有一种简便方法，即“中位数”。中位数是指把所有数据按从大到小（或从小到大）的顺序进行排列时，排在最中间的数据。顾名思义，中位数就是位于中间位置的数据。有奇数个数据时，中位数是正中间的数据，有偶数个数据时，中位数则是中间两个数据的平均值。

例如，如果有 4 个数据（0、3、6、9），那么中位数就是 3 与 6 的平均值 4.5。中位数可以用 Excel 函数直接得出来，记住这个函数和平均值的函数会很方便。

•平均值：=AVERAGE（在此处指定数据范围）

•中位数：=MEDIAN（在此处指定数据范围）

如果以平均值为中心，所有数据都大致均等地分布在其左右两侧，那么中位数的值必然会接近平均值。相反，如果受离群值的影响，平均值在所有数据中的位置较偏，那么中位数与平均值之间就会出现较大差异（也有个别情况例外）。我们也可以在这种情况下再用直方图等查看平均值为什么会出现偏离。

图 2-10 为平均值和中位数的特征一览表。详细了解平均值和中间值的各自特征，并根据需要组合运用，可以避免落入滥用平均值的陷阱。

图 2-10 平均值与中位数的特征

	优 点	缺 点
平均值	• 能够用一个数值表现整体的"大小"（不一定位于中间位置） • 易于计算及使用 • 认知度高，便于与其他人沟通	• 平均之后，看不出原始数据 • 存在极大（或极小）数值时，会受其影响
中位数	• 不受离群值影响，用位于中间的数据表示 • 易于理解，便于与其他人沟通	• 与平均值相比，表示整体"大小"的程度不够精确（这是由于消除了离群值影响而产生的缺陷） • 无法通过计算得出（可以利用 Excel 的 MEDIAN 函数得出） • 与平均值相同，看不到原始数据

"油炸豆腐"和"天妇罗面渣"，哪个更好吃

根据平均值进行比较，不仅要注意原始数据，还需要注意"数据构成"的差异。

例如，假设您是一家荞麦面馆的老板。为了了解客人喜欢的配菜是"油炸豆腐"还是"天妇罗面渣"，分别向 200 位客人做了问卷调查，请他们打分（满分 100 分）。其结果如图 2-11 所示。

图 2-11 对油炸豆腐和天妇罗面渣的评价

	油炸豆腐	天妇罗面渣
回答人数	200	200
总平均分	67.0	60.8

大概很多人会认为：油炸豆腐的（总平均分）高出 6 分多，那就多进些油炸豆腐来卖吧。

不过在店里打工的学生说："我们店里既有乌冬面，也有荞麦面，所以我对哪一种更受欢迎做了统计"，并拿来了结果。如图 2-12 所示，他分别统计了乌冬面和荞麦面的平均分。

图 2-12 对乌冬面与荞麦面的评价

平均分	油炸豆腐	天妇罗面渣
乌冬面	70	80
荞麦面	40	45

令人吃惊的是，无论乌冬面还是荞麦面，都是天妇罗面渣的平均分更高。他的表格没有任何计算错误或数据选取错误，只是正常计算了平均值，但结果却与图 2-11 截然相反。如果没有这名学生提出疑问，您可能会根据最初的结论，判断是油炸豆腐更受欢迎。

这种矛盾的现象被称为"辛普森悖论"，可能很多人都不太了解。有些情况下，考察数据整体和考察不同部分，会得到相反的结论。如果您是店老板，对这种情况会做何判断呢？

平均值所掩盖的真相

上述过程中隐含着回答者构成的不同。此次问卷调查的回答者人数分布如图 2-13 所示。

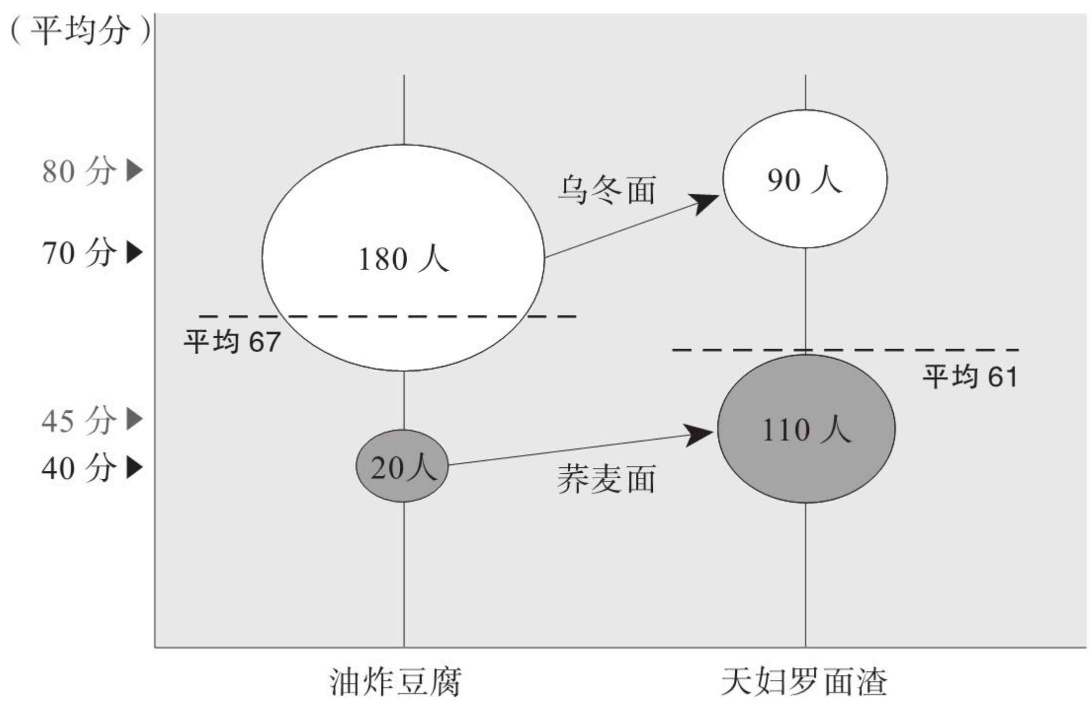
图 2-13 油炸豆腐与天妇罗面渣、乌冬面与荞麦面的回答人数分布

回答者人数	油炸豆腐	天妇罗面渣
乌冬面	180	90
荞麦面	20	110
合计	200	200

也就是说，为油炸豆腐打分的大部分回答者是吃乌冬面的人，而为天妇罗面渣打分的回答者则多是吃的荞麦面。简单地说，是回答者在“乌冬面”和“荞麦面”之间的不平衡分布影响了总平均分的结果。

从图 2-12 可知，乌冬面的分数高于荞麦面。也就是说，与荞麦面相比，乌冬面会拉高分数。选择“油炸豆腐”的人中，吃乌冬面的人数（180 人）远远多于吃荞麦面的人数（20 人），乌冬面拉高了“油炸豆腐”的平均分。看到这里，可能很多人脑子里会画上一个“？”，那么我们就再从直观上确认一下（图 2-14）。

图 2-14 油炸豆腐与天妇罗面渣出现不同比较结果的玄机



•乌冬面的分数 > 荞麦面的分数

•油炸豆腐：吃乌冬面的人数 > 吃荞麦面的人数

•天妇罗面渣：吃荞麦面的人数 > 吃乌冬面的人数

这三个不等式同时成立，就产生了“油炸豆腐的平均分高于天妇罗面渣”的结果。但分别看乌冬面和荞麦面的评分结果，却都是“天妇罗面渣的平均分高于油炸豆腐”。

乌冬面和荞麦面的教训

只看整体，我们可能注意不到“数据构成要素的差异（这里是乌冬面与荞麦面的不同）”，忽略这种差异进行单纯比较，就有可能导致无法察觉该差异所造成的影响。正如这个事例所显示的，关注原始数据内部的要素，有时会看到不同的风景。

在这个事例中，通过留意乌冬面与荞麦面的要素，我也许会转为专注乌冬面与荞麦面的差异。也就是说，问题并不是油炸豆腐和天妇罗面渣的不同。

一般而言，将数据细分至何种程度才能做出适当判断，这个问题并没有标准答案。为了找出正确答案，分析者需要了解自己的数据可以分为哪些层次，目前分析处于何种状态，这一点很重要。此外，进行比较时，了解样本的选取方法也很重要。

作为平均值的特征，我们应该了解：平均值掩盖了“乌冬面与荞麦面回答人数的不同”。

事先了解到平均值虽然方便，但也具有这样的副作用，我们就可以在有意识的前提下面对平均值。

我平时看到平均值，会更为关注“眼前这个数值是哪些要素综合计算出来的”。

如果能注意到一笔笼统的账目中到底包含了什么，或许就可以像这个事例一样，对乌冬面与荞麦面的要素进行分解，从而找到其中隐藏的信息。也就是说，平均值的背后可能还蕴含着很多沉睡的宝藏。

用“波动”的视点给平均值做补充

尽管需要留意，但平均值仍然是无人不晓的便利工具。如果有人问“工作中哪些场合会用到平均值”，我会回答：“想大致了解对象大小的时候”。当然，我还会补充一句，“同时也要关注平均值背后的隐秘信息”。我们可以记住以下两个重要事项。

①平均值最适合表示整体的大小

平均值便于人们把握大体趋势，可以通过平均单价等把握和比较整体情况。

②平均值的缺点是会掩盖原始数据的部分特征

注意到这一点，可以避免遗漏更为精确的信息。

只从大小的维度处理数据，会遗漏许多信息，但实际工作中的绝大部分资料都只有平均值，完全不提及数据的波动。

请看图 2-15 的例子。如果只按平均销售额（粗框内）来看这家店铺的业绩，会得出什么结论呢？因为从 6 月到 7 月期间平均销售额没有变化，所以结论一定会是“没有什么变化”吧。也就是说，结论是无须过多担心。

图 2-15 某店铺 6 月与 7 月的销售额比较

业绩	日平均销售额（万日元）	每日销售额的波动
6 月	45	小
7 月	45	大

但是，如果我们将关注点从日平均销售额转到相关月份里每日销售额出现波动的程度，也许就会发现，6 月到 7 月出现了每日销售额波动幅度增大的情况。

在商业世界中，波动也可以说是“风险”。虽然最初得出“没有变化”的结论，但它很可能是错误的，正确的结论应该是“风险增大了”。只注意平均值的人忽略了这一点，因此无法得到正确的结论。

需要注意的是，“波动大”或“风险大”并不一定就是坏事。风险也可以反过来看作机会。有时我们也可以有意选择“高风险、高回报”的情况。

不过，对希望实现精确的库存管理、希望每日销售额尽量保持稳定的人来说，体现波动（风险）增大的指标对店铺管理具有重要意义，是极其重要的信息。对于如此重要的信息，决不能只是漫不经心地浏览一下数据，就主观地判断是“大”还是“小”。更何况仅根据最近的数据变动情况，也未必能够掌握准确的数据特征。

能够客观体现波动大小的指标叫作标准差。标准差表示以平均值为中心，数据对平均值的偏离程度。人们可以根据标准差的数值对类似每日销售额的波动程度等进行相对评价。

之所以要做相对评价，是因为只看标准差的大小，很难得出有价值的信息。只有在相同条件下（例如对同一家店铺的不同月份进行比较、对销售额规模相同的不同商品进行比较等），比较相对大小才能发挥作用。

如图 2-16 所示，对同一家店铺的不同月份进行比较，虽然平均值都是 45 万日元，但由于标准差从 10.1 万日元增至 20 万日元，所以可

以定量地体现出风险的增加。为了避免类似“感觉数字的浮动幅度似乎有所增大”等含糊不清的意见，或者由于数据太多，连这种直觉都无法发现问题时，可以用标准差的指标客观地进行判断。

图 2-16 平均值相同，标准差可能不同

	6 月	7 月
日销售额（万日元）	48	22
	46	35
	39	47
	27	57
	35	69
	38	28
	40	73
	47	61
	36	79
	56	33
	41	78
	55	22
	42	72
	50	28
平均值	45.0	45.0
标准差	10.1	20.0

这样一来，就可以进一步研究和讨论“为什么 7 月份销售额波动会增大”。找到原因，才能根据需要采取措施。

不过在实际业务中，标准差会受到很多制约，不像平均值那么易于使用。这也是个令人头疼的问题。

例如，个体商店与百货商场的日销售额相距甚远，比较二者的标准差不具有任何意义。因为原始数据的大小不同。

日平均销售额 1 000 万日元的店铺波动幅度（标准差）是 30 万日元，日平均销售额 100 万日元的店铺也有 30 万日元波动，如果说二者“波动幅度相同”，恐怕谁都会觉得不对头。

此外，很多书都从统计学的角度介绍标准差，其代表性的说明如下：

“以平均值为中心，在向其左右各扩大 1 个标准差的范围内，会包含约占整体三分之二的数据（满足正态分布的条件下）”

第一次看到这句话的读者，恐怕很难立刻理解其含意吧。这句话换成更易于理解的说法，就是“在所有数据以平均值为中心呈左右对称分布、数据个数随着数值对平均值的偏离而减少的情况下，在平均值增加和减少 1 个标准差的范围内，会包含所有数据的三分之二（如果有 100 个数据，就会有约 66 个数据分布在平均值增加和减少 1 个标准差的范围内）”。即使这样说，大概还是会有很多人不明白吧。

更麻烦的是，即便理解了这段话的含义，在大多数情况下，“包含三分之二数据的范围”也很难直接应用到实际业务中。

“以平均值为中心左右对称”“数据集中在平均值附近，极端数据较少。如果不符合这些（正态分布的）前提，标准差就没有太大意义。然而我们很难断定所有数据都接近正态分布。不如说在商业领域里，相反的情形更多。

标准差为什么“用不上”

因此，很多人因为“统计书上的内容”与“自己运用”之间存在明显的距离（还有很多情况下，二者之间的距离甚至大得已经感觉不到距离），而放弃使用标准差。

这就是人们常会觉得“理论上好像明白，但不会实际运用”标准差这个工具的原因。

标准差的计算公式如图 2-17 所示。为了去除正负影响，先把各个数据的值与平均值的差平方后相加，除以数据个数之后，再开方。简而言之，就是用一个数值来表示各数据与平均值距离的和。

图 2-17 标准差的计算公式

$$\text{标准差} = \sqrt{\frac{（\text{各数据值} - \text{平均值}）^2 \text{ 的和}}{\text{数据个数}}}$$

用计算器计算标准差很麻烦，可以使用 Excel 函数立即得到结果（图 2-18）。此外，不同版本的 Excel 中还有很多类似的函数。

图 2-18 Excel 中的标准差函数定义

= STDEV（在此处指定数据范围）

当然，只要不是严谨的学术领域，我们在现实工作中也可以不考虑其前提条件，而只是粗略地比较数据的波动程度。但重要的是，如果我们在自己完全不了解（或者是忽略了）原本的制约条件下得出结果，那么在运用该结果时应该知道它并不是严密的。

我在日产工作时曾经制作了一个系统，用来管理遍布世界各地的 120 个国家或地区的销售代理店业绩。在数十项评价指标当中，有一些无法只靠绝对值评价好坏，也有的指标需要在所有代理店中进行相对评

价。当时我采用的方法是，将平均值增减 1 个标准差的范围定义为“标准”范围，关注处于该范围之外的数据（代理店）。如果随便设定一个标准，据此来评价好坏的话，人们有可能会质疑评价本身是否合理，但如果能够说明客观标准是如何设定的，就不太会出现类似问题。

作为标准差的应用事例之一，再来介绍一下学校里常会用到的偏差值。步入社会以后，除非在教育领域工作，否则可能很少有人会用到偏差值，但这里还是做一简单介绍，来加深大家对标准差的理解（图 2-19）。

图 2-19 偏差值的计算公式

$$\text{偏差值} = \frac{10 \times (\text{得分} - \text{平均分})}{\text{标准差}} + 50$$

例如，如果有人在平均分为 50 分的考试中得了 80 分，虽然知道这个分数“高于平均分”，但却无法得知这个人在整体中的位置。80 分与平均值之间的 30 分差距是“大”还是“小”，取决于所有分数的分布情况。

如果得分超过 80 分，甚至更高的人数很多（即标准差较大），就不能说 80 分是一个相对优秀的结果。而如果大部分人的分数都在平均分 50 分前后（即标准差较小），则可以说 80 分是“非常优秀”的成绩。偏差值就是用来表示“成绩得分的相对位置”的指标。

那么，有没有什么办法可以将不太好用的标准差运用到实际工作中呢？这里介绍两种能够轻松掌握数据分布的方法。

一种方法是用柱状图来展现数据的分布情况，可从视觉上直接确认。在前文提及的直方图中，纵轴表示数据的个数（频数、频率），横轴表示数据大小的间隔（区间）。经常有人会问间隔的“最佳幅度”是多少，这个问题并没有标准答案。在实际业务中，一般由分析者根据目的或者是否便于理解来自行决定。

比较两个直方图，可以看出哪一边的数据波动更大（或更小）。使用直方图将全部数据直观体现出来，可以帮助人们把握数据分布的形状、特定范围内的数据个数等只看标准差所无法获得的信息。

与在 Excel 中直接查看数据相比，做成直方图可以更轻松地把握实际情况。直方图只是从直观上把握数据，所以不必在意计算标准差的前提以及评价标准差的方法等比较难懂的细节问题。因此，向其他人展示或者进行说明时，直方图的魅力之一就是可以作为更便于理解的信息发挥作用。

直方图也是我在日产工作时经常用到的方法之一。如果不了解对方的专业水平或关注点所在，或者要向来自不同国籍或部门的高管进行展示时，这是一种更为“保险”的方法。虽然做直方图需要使用大量数据，但其优势是任何人都能很快理解。当然，这种方法在数据的特征和结果较为明显时效果会更好。而对于很难从直观上判断的情况，使用直方图反而会招致混乱。

“变异系数”的魅力

还有一种方法是变异系数。使用标准差对波动程度进行相对比较时，必须满足被比较的两个数据大小相当或相同（例如同一家店铺不同月份之间的比较、相同销售额规模的两家店铺之间的比较等）的前提条件。因为一般情况下，如果原始数据值较大，其波动（标准差）也会相对比较大。不消除“数据大小”的差异，就无法进行适当的比较。

说到数据的大小,请各位回忆一下平均值的相关内容。如果能用标准差除以全部数据的平均值，消除数据大小的差异，就可以不必在意原始数据的大小，从数值上用标准差来比较两者的波动程度。标准差除以平均值得到的值叫作变异系数。

顺便提一下，我们无法单独评价变异系数，这个指标只能用来比较不同数据的波动程度。例如，图 2-20 为大规模店铺与小规模店铺的日销售额。只看标准差的话，大规模店铺的波动更大，为 50。

然而考虑到二者之间原本存在销售额的差异，不能这样做单纯比较。计算变异系数，则可以消除销售额规模的差异，从而得知小规模店铺的数据波动更大。也就是说，小规模店的经营风险相对更大。

图 2-20 大规模店铺与小规模店铺的比较

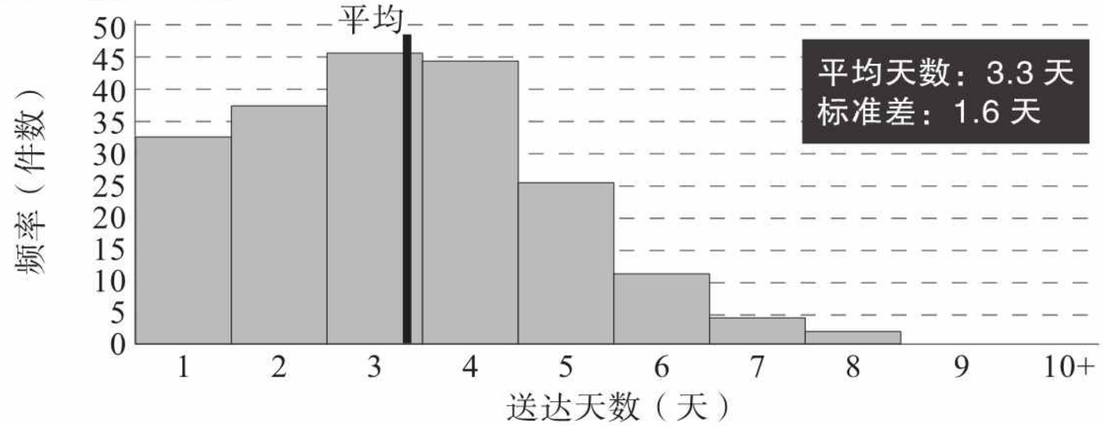
	大规模店铺	小规模店铺
平均销售额（万日元）	1500	175
标准差	50	25
变异系数	0.03	0.14

变异系数的魅力在于计算简单，而且与其他统计指标一样，可以只用这一个指标进行客观比较。有些情况下，它还可以弥补直方图的如下缺陷。

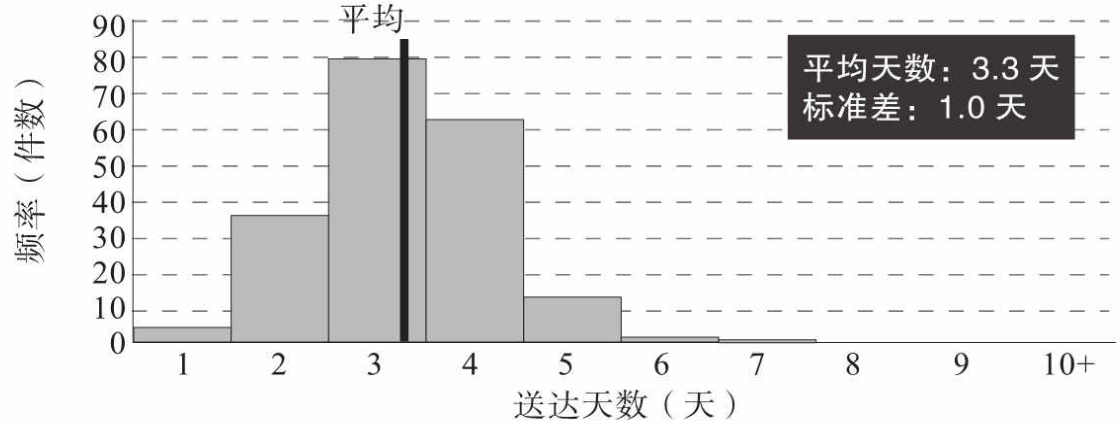
例如，记录 A、B、C 三个区域的快递送达天数，制成直方图（图 2-21）。

图 2-21 三个区域送达天数的不同

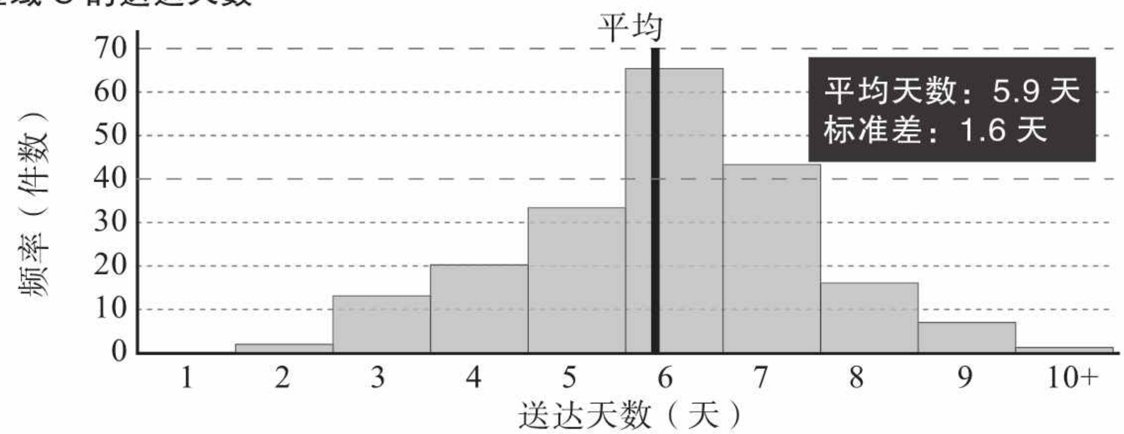
区域 A 的送达天数



区域 B 的送达天数



区域 C 的送达天数



	区域 A	区域 B	区域 C
平均值	3.3	3.3	5.9
标准差	1.6	1.0	1.6
变异系数	0.5	0.3	0.3

如果根据这些直方图来讨论哪个区域送达天数的波动更大（注意讨论的不是送达天数），人们的意见可能会出现分歧。事实上，由于区域C的数据向左右扩展的范围看似更大，所以很多人会认为该区域的波动较大。我在培训中展示这个直方图时，大家会根据表面现象做出不同的主观判断。这样就无法对讨论的前提形成一致。

视觉也是“感觉”

所谓视觉，也是一种“感觉”，具有容易让人根据主观进行判断的一面。可以说最能在这种场合发挥威力的，正是根据数据做出的客观判断。

图 2-21 中还标出了每个区域的平均送达天数和标准差。在这个事例中，可能是因为只有区域 C 的覆盖范围较广，所以其平均送达天数较长，为 5.9 天。

如果仅比较平均送达天数，就会得出“区域 C 波动幅度大”的结论，毫无疑问这是错误的。不过前面已经说过，这种情况也不适合只用标准差进行比较。

计算各区域的变异系数，区域 A 最大，为 0.5，所以结论是区域 A 波动幅度相对最大。像这样，用数值进行比较的结果一目了然，不会因为人们对“表面现象”的主观判断差异而产生分歧。这是在组织中进行决策的一个极其重要的研究方法。不过正如前文所述，我们需要先考虑“由此能够获知什么”“它在这项问题中具有何种作用”，然后再计算变异系数。如果只是在形式上计算出变异系数，对其含义却并不理解的话，就是本末倒置了。

从易于理解和便于接受这一点来看，直方图非常好用，但其问题是在 Excel 中无法用一个操作直接制作（市面上销售的一些统计软件具有

这个功能）。因此需要我们根据实际数据，设定适当的区间间隔，计算各范围内的数据个数，在此基础上才能做成图表，所以略为费事。

图 2-22 原始数据与数据区间（分组）

	A	B	C
1	原始数据	数据区间（分组）	
2	81	10	~10
3	49	20	11~20
4	63	30	21~30
5	68	40	31~40
6	59	50	41~50
7	32	60	51~60
8	47	70	61~70
9	42	80	71~80
10	74	90	81~90
11	62	100	91~100
12	49		
13	32		
14	50		
15	92		
16	72		
17	26		
18	16		
19	70		
20	55		

在此简单介绍一下用 Excel 制作直方图的方法。对于平时不常使用 Excel 的人来说，首先需要启动 Excel“加载项”菜单中的“数据分析”。

假设 A 列为原始数据（图 2-22）。在另外一列（此处为 B 列）输入数据间隔。这里按照 10、20、30……100，以 10 为间隔来输入。它表示直方图中区间幅度的最大值。为了便于理解，也为了以后用作直方图的横轴，在 C 列手动输入每个区间的具体范围。

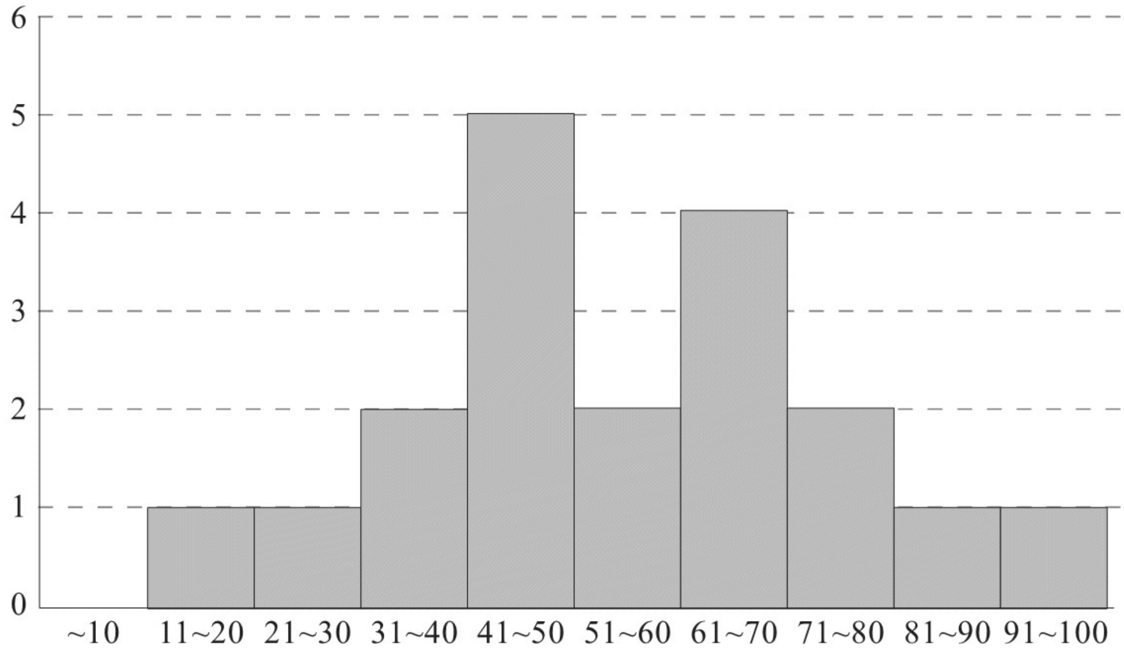
然后在加载项中启动“数据分析”，选择“直方图”。在“输入区域”指定数据范围，在“接受区域”指定自己设置并输入数据间距的范围。然后选择便于自己查看的区域输出结果。按下“确定”键之后，就会如图 2-23 所示，在 E 列和 F 列显示出各数据区间及其对应的数据个数（频率分布表）。

图 2-23 每个数据区间的数据数量（频率）

D	E	F	
	数据区间	频率	
	10	0	
	20	1	
	30	1	
	40	2	
	50	5	
	60	2	
	70	4	
	80	2	
	90	1	
	100	1	
	其他	0	

接下来，将上述“频率”做成柱状图。直方图中的柱体一般相互连在一起，没有柱状图中的间隔，需要自己调整。对任意一个柱体点击右键，打开“设置数据系列格式”，将“分类间距”改为“0”。这样，就完成了一幅如图 2-24 所示的直方图（Excel2013“加载项”的“直方图”中包含制作图表功能）。

图 2-24 完成后的直方图

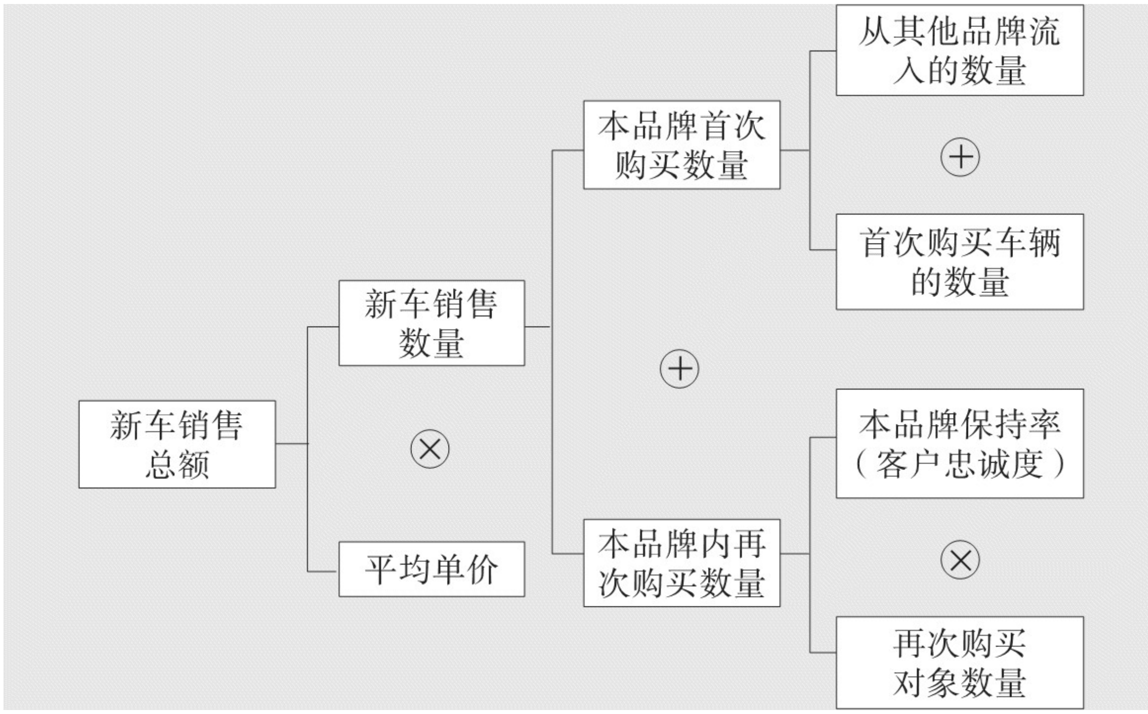


「解决问题的故事 2」

第二步：“分解”数据，锁定问题的关键

对问题做出恰当的定义，并大致把握现状之后，接下来需要找到“问题的关键”。通过找出问题的关键，可以实现确定分析范围的目的。如果数据中包含所有信息，就会像一笔糊涂账，一直钻研也无法取得任何进展。因此，分解数据就显得极为重要。

图 A 分解新车销售总额



如图 A 所示，“新车销售总额”可以通过四则运算，用“WHAT 型假设”进行分解。左端的“新车销售总额”是代表最根本问题的指标，可以分解为“新车销售数量”与“平均单价”的乘法。

分解“销售总额”的构成要素，可以使分析者看到更为具体的内容。从新车销售数量和平均单价两个方面，可以比只看销售额挖掘得更为深入。当然，在这种假设的基础上进行分析，也还是有可能无法找到任何值得参考的信息。通过数据分析可以对此加以判别。

很多情况下，如果进展不够顺利，就得反复试错，回到假设阶段寻找其他突破口。我们需要放弃尝试一次便获得成功的完美主义。

仅根据新车销售数量的变化，还远远无法找出“问题的关键”。因此，还需要将新车销售数量分解为“本品牌首次购买数量”与“本品牌内再次购买数量”的加法。也就是说，将首次购买本品牌汽车的用户与再次购买本品牌汽车的用户加以区分。

此时可以假设这两种用户之间存在不同的购买动机。很多情况下，人们根据常识就可以建立类似假设，了解相关业务的人，做出

错误假设的可能性也不会太大。

如果认为进一步分解能体现出数据特征的差异，可以继续分解（虽然比较麻烦），从而提高分析的准确度。在这个例子中，我们尝试进行了更深入的挖掘。

图 B 根据图 A 进行数据整理的结果

新车销售总额（万日元）		新车销售数量（辆）	平均单价（万日元）	本品牌首次购买数量（辆）	本品牌内再次购买数量（辆）	再次购买对象数量（辆）	客户忠诚度（%）
1 月	36 605	189	194	89	100	124	80.4
2 月	36 912	181	204	95	86	121	71.0
3 月	36 437	181	201	91	90	130	69.4
4 月	36 667	189	194	84	105	134	78.4
5 月	36 317	176	206	87	89	135	66.1
6 月	36 553	187	195	87	100	123	81.7
7 月	36 794	198	186	82	116	139	83.3
8 月	35 883	178	202	85	93	134	69.1
9 月	35 846	182	197	95	87	124	70.1
10 月	36 052	189	191	86	103	125	82.2
11 月	35 754	172	208	89	83	132	62.8
12 月	35 029	168	209	82	86	132	64.9
1 月	35 480	174	204	91	83	124	66.9
2 月	34 854	168	208	94	74	120	61.3
3 月	34 725	181	192	93	88	128	68.6
4 月	33 180	172	193	83	89	130	68.4
5 月	33 592	170	198	90	80	123	64.8
6 月	32 538	162	201	93	69	128	53.8
7 月	32 545	156	208	88	68	124	55.2
8 月	32 462	165	197	84	81	120	67.3
9 月	32 172	161	200	92	69	125	55.1
10 月	31 855	164	194	93	71	130	54.8
11 月	31 656	154	205	87	67	128	52.7
12 月	30 940	153	202	82	71	120	59.3

我们将“本品牌首次购买数量”，即首次购买本品牌产品的顾客分为“从其品牌流入”的用户和首次购车（非再次购买）的用户。

接下来，我们将“曾经买过本品牌产品、现在需要再次购买的用户”分解为再次购买时仍然选择了本品牌产品的用户和转为购买其他品牌产品的用户。我们将重新购买新车的本品牌用户中，再次选择了本品牌产品的人（回头客）的比例作为“客户忠诚度”。

图 B 对这些内容做了数据整理。利用 WHAT 型假设进行数据分解，在结构上具有一个优点，即能够明确分解后的数据（例如“客户忠诚度”“从其他品牌流入的数量”等）与上一级变量（“新车销售数量”）以及最上层问题（“新车销售总额”）之间的关系。

掌握了分解之后的各个项目的大致情况（倾向），我们便可以找到更值得关注的关键。

从图 C 可以发现，新车销售总额在这 2 年期间减少了大约 15%。这是所有问题的根源。根据图 D 可知，平均单价在 2 年期间比较稳定，基本在平均值（200 万日元）上下 5%（190 万～210 万日元）的范围内。至少在过去的一年里，没有出现过价格明显下降的情况。另一方面，从图 E 可以看出，“新车销售数量”与销售总额一样，也减少了近 15%。由此可知，我们应该优先关注新车销售数量的数据。接下来，我们可以进一步观察图 F 所示的销售数量的明细。

图 C 新车销售总额的变化

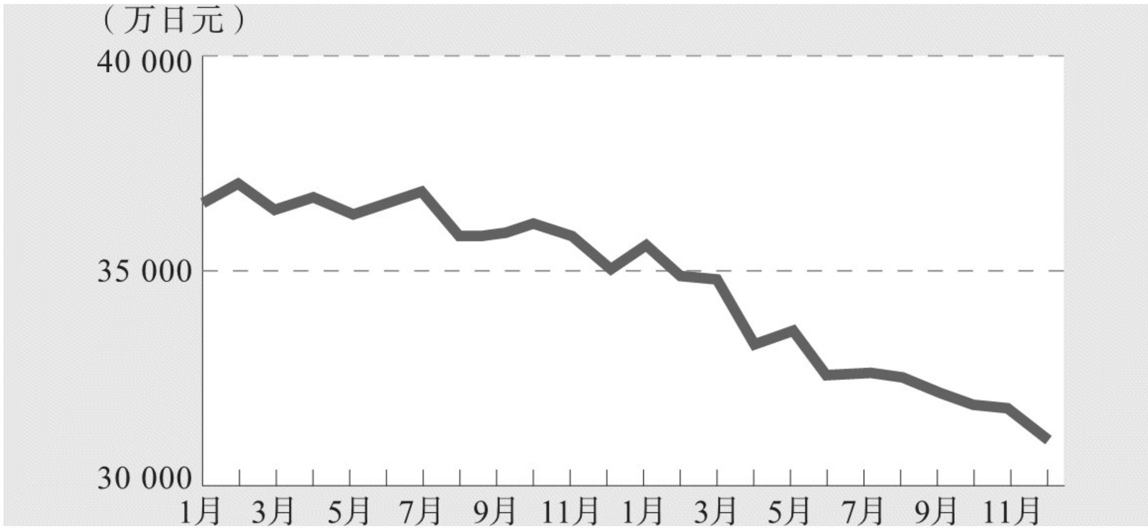
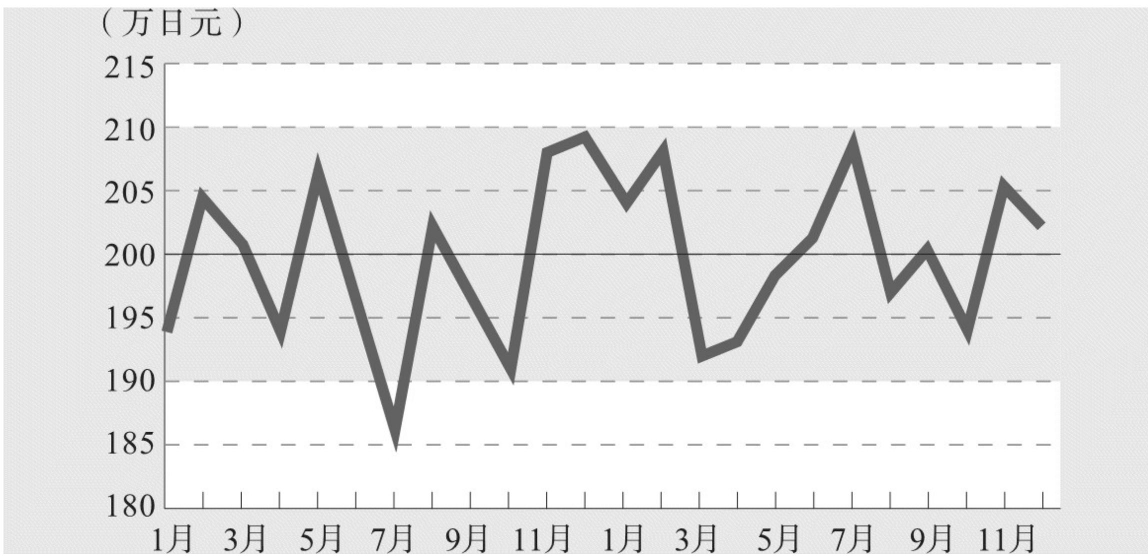


图 D 平均单价的变化



虽然两者在数量上没有太大差异，但本品牌首次购买数量基本维持稳定，而再次购买数量却在过去一年出现了减少。因此，从WHAT 型假设的构成也可以确定，“本品牌内再次购买数量”应该是“新车销售总额”减少的主要

图 E 新车销售数量的变化

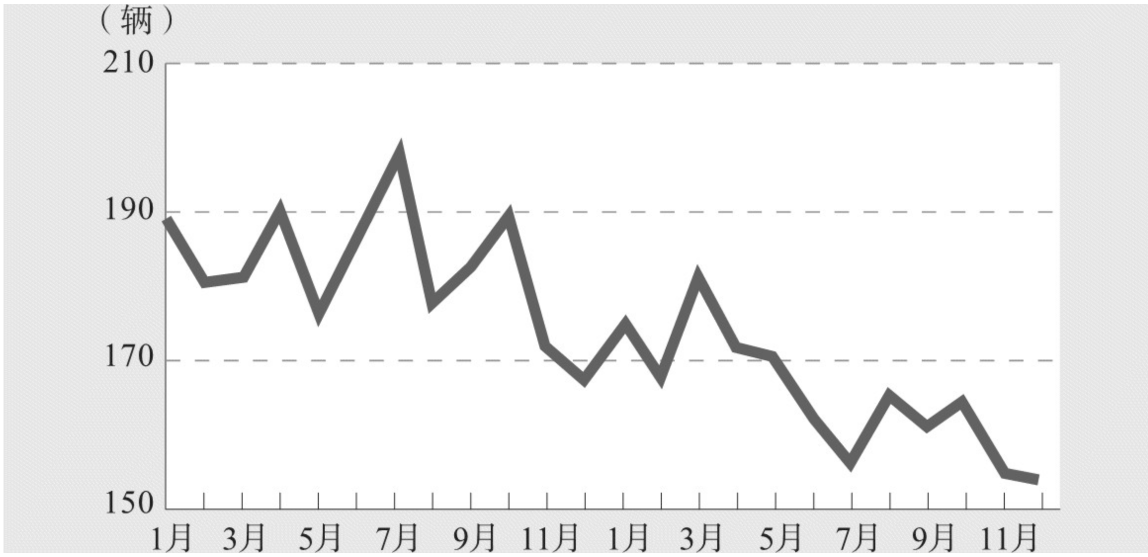
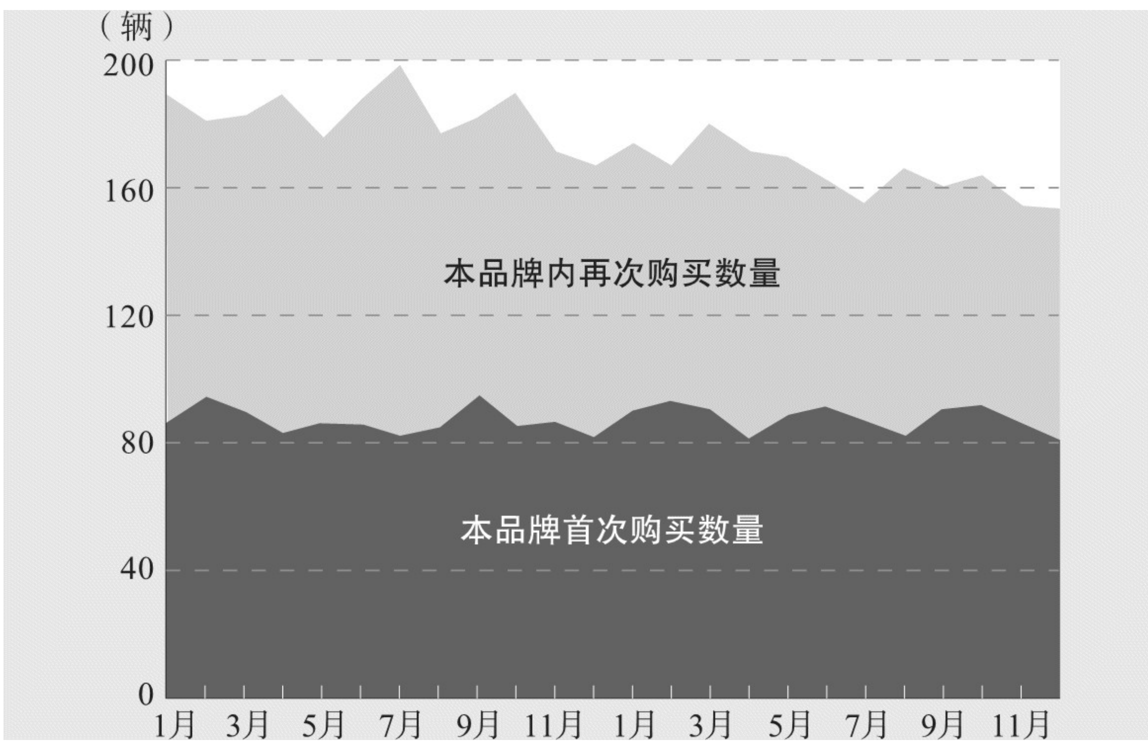


图 F 新车销售数量明细的变化



原因之一。不分析到这一步，即使发现“新车销售额（数量）减少”，也仍旧无法回答“是否存在其他问题”“如何断定这就是主要问题”等质疑或追问。这样的话就难以保证分析的可靠性。

要注意的是，不能根据“本品牌内再次购买数量”本身来判断“数量增加就好”或者“数量减少所以不好”。我们应该意识到，根据单纯的数量“增加或者减少”“多或者少”只能看到表面的结果，而很难对现状做出恰当的评价。

因为只有将“本品牌内再次购买数量”与当月再次购买车辆的总人数（辆数）进行比较，才有可能对现状做出评价。为此，需要进一步分解，掌握在所有需要再次购买的数量当中，有百分之几的人实际再次购买了本品牌产品。

在看到某个数字时，我们要关注的是，“这个数字能否真正评价我们想知道的问题”“这个数字是在哪个范围产生的”“在与其他数据进行比较时，是否需要统一标准（这个例子中的标准是每月再次购买车辆的人数）”等问题。

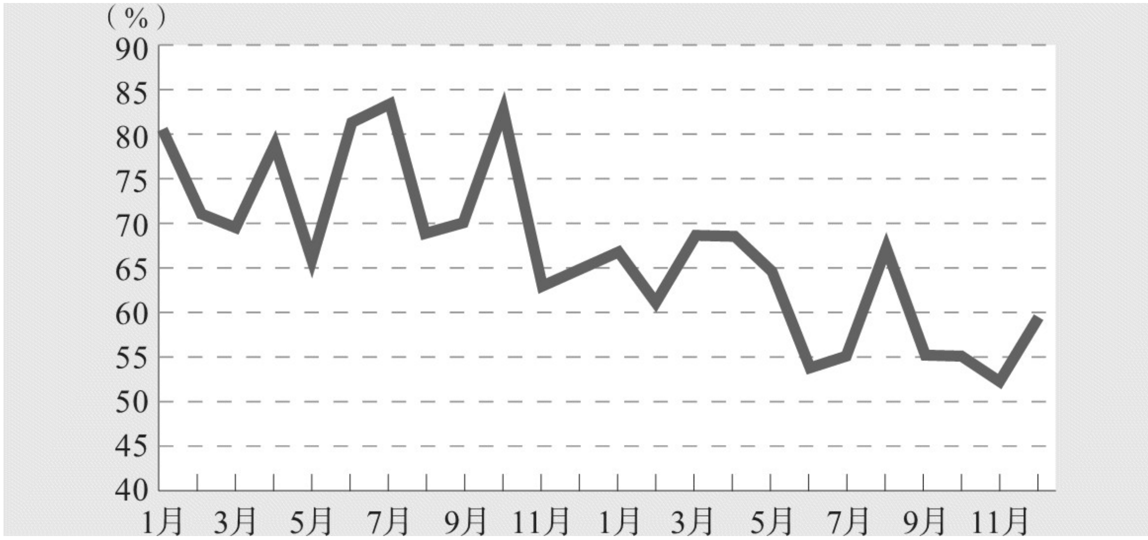
将“再次购买数量”分为“本公司品牌”和“其他公司品牌”，并不只是因为可以从理论上这样做。之所以这样分解，是因为如果能确定这其中存在问题，就可以采取具体的措施。了解再次购买汽车时转为购买其他公司产品的人与再次购买本公司产品的人的不同，才能采取对策。衡量这一情况的不是具体数量，而是客户忠诚度（%）。

只有将业务经验、常识及其他人的意见都动员起来，才能获得创意和思路。因为数据自己是不会讲话的。

反之，如果一个数据分析没有经历过类似的思考过程，就可以说它没有目的和终点。实际上，到处都可以看到这种“机械性”的数据分析。

观察客户忠诚度的变化（趋势），可以发现导致“本品牌内再次购买数量”减少的，不是整体再次购买数量的减少，而是客户忠诚度的下降。也就是说，分流至其他品牌的人数的比例增加了（图G）。

图G客户忠诚度的变化



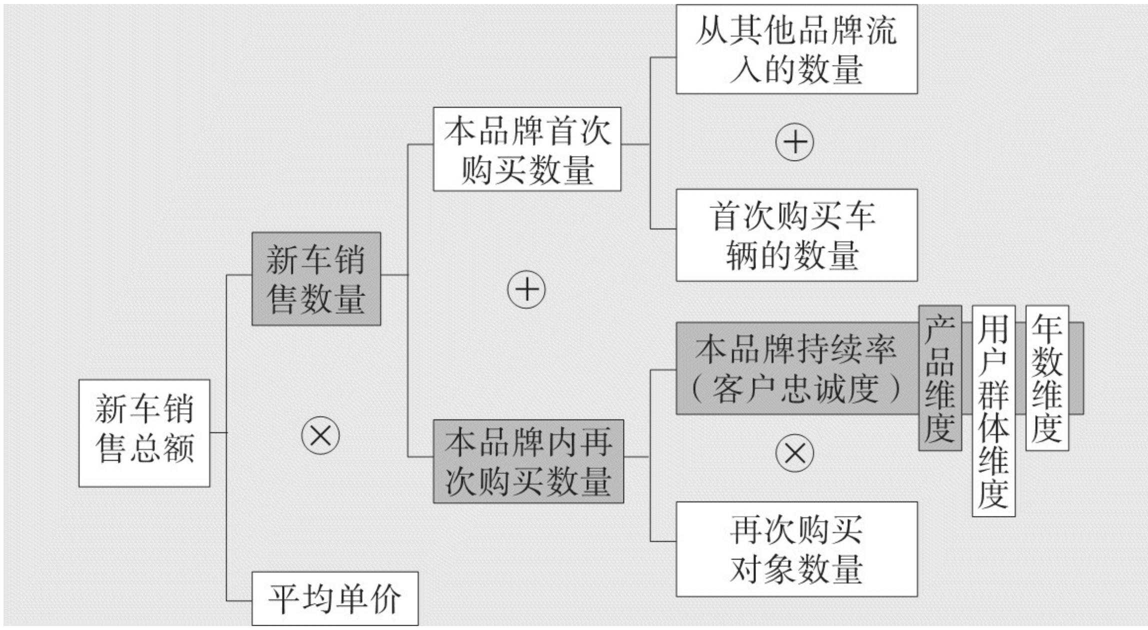
对于“本品牌首次购买数量”，本来也需要用相同的方法把握整体情况。但因为无法通过一本书覆盖所有的内容，所以本书仅针对现状中已经发现的“客户忠诚度”的问题，即如何让用户再次购买车辆时仍然选择本品牌产品这一问题做进一步的深入挖掘。

比较的维度

虽然发现了客户忠诚度的问题，也还无法就此锁定问题的关键。同样是客户忠诚度，在按照某个维度比较时，还有可能会包含与问题有关的要素和无关的要素。用户群体（年龄、职业、性别等）、年数、车型（产品）等，从哪个维度来比较客户忠诚度，才能锁定问题的关键呢？决定维度之前的所有工作，都是建立WHAT型假设的一部分。作为这个例子的可靠假设，可以展开以下讨论。

按照不同的客户群体，例如不同年龄层的客户来分析客户忠诚度，会发现对不同的车型来说，人们的评价或喜爱

图H 突出体现假设的要素（带灰色背景的项目成为问题的关键）



程度并不一定与年龄有关。也就是说，不同的车型更能体现出人们对其评价或喜爱程度的不同（图 H）。

根据这一假设，可以从“产品（车型）”的维度来比较不同产品的客户忠诚度。首先，我们来整理一下应该怎样比较（图 I）。

图 I 从产品维度进行比较

	“快照”的视点	“趋势”的视点
大小或比例的维度	（1）对一定时期内的新车销售数量和客户忠诚度的平均值进行比较	（2）对客户忠诚度的变化进行比较
波动的维度	（3）对一定时期内客户忠诚度的波动进行比较	（非此次检验对象）

（1）对新车销售数量和客户忠诚度在一定时期内的平均值进行比较

对销售总额这个最根本的问题来说，是否存在对其影响较大的车型和并无太大影响的车型呢？虽然深入挖掘是好事，但如果挖掘

的对象对问题整体的影响微不足道，这个工作就不会产生任何意义。根据对问题影响的大小来决定优先顺序和关注程度，也决定了根据分析结果采取的措施能够对解决问题产生多大贡献。

对销售数量的构成比例进行比较，可以发现只有车型 A 的比例明显小于其他车型（图 J）。那么如果将问题锁定为车型 A，即使采取了有效的对策，对解决整体问题的影响仍然是有限的。因此可以暂且降低车型 A 的优先顺序。

此外，按照不同车型，对客户忠诚度在 2 年期间的平均值进行比较，只有车型 A 的客户忠诚度显著偏低，其他车型之间没有太大差别（图 K）。也就是说，从“快照”的视点无法发现车型 B 到车型 D 之间的差异。读者们肯定也会感到“怎么会这样？”

图 J 不同车型在新车销售数量中的比例

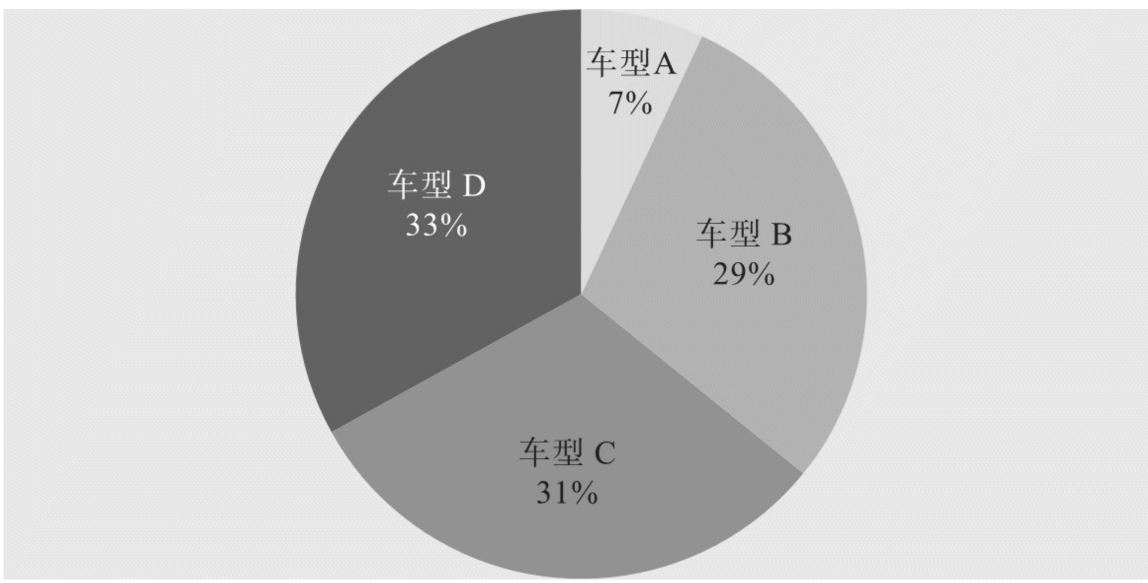


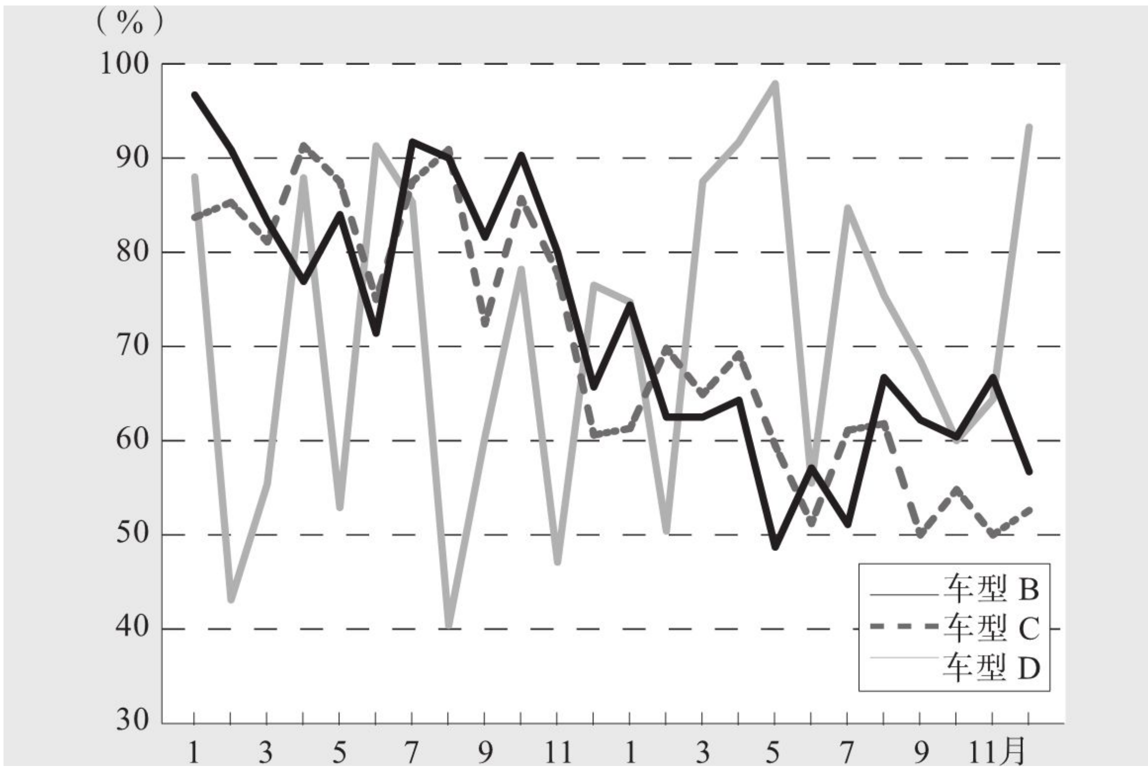
图 K 不同车型的平均客户忠诚度

车型 A	车型 B	车型 C	车型 D
48.9%	72.3%	70.2%	71.3%

（2）对客户忠诚度的变化进行比较

接下来，利用折线图对过去 2 年期间不同车型的客户忠诚度进行比较。根据前面确认的结果，暂且将车型 A 从比较对象中剔除，对其余 3 个车型进行比较（图 L）。

图 L 不同车型的客户忠诚度的变化



这时可以发现，车型 B 和车型 C 的客户忠诚度均自一年前开始逐渐降低。也就是说，转为购买其他公司产品的比例提高了。具体数字是 2 年期间从约 80%～90% 减至 50%～60%，降低了 30～40 个百分点。可见，拥有本品牌产品但需要再次购买车辆的人中，这 2 年期间有近 30% 被其他公司夺走。

（3）对一定时期内客户忠诚度的波动进行比较

接下来对 2 年期间客户忠诚度的变异系数进行比较（图 M）。

图 M 不同车型客户忠诚度在 2 年期间的变异系数

车型 A	车型 B	车型 C	车型 D
46.0%	19.3%	20.0%	24.9%

虽然车型 A 的变异系数明显高于其他车型，但该车型数量较少，对整体的影响很小，因此在此就不涉及。从其他车型来看，正如前面的客户忠诚度的变化（图 L）所示，车型 D 的波动较大。从平均来看，车型 D 的客户忠诚度与车型 B、车型 C 几乎没有差别，在 2 年期间也并未出现恶化，所以很难将其认定为直接影响问题的关键。但从波动程度来看，车型 D 极有可能含有其他问题或风险，需要引起注意。

当数据的数量更多或者期间更长时，从趋势的视点来观察波动的变化也会发挥一定作用，但因为这里只有 2 年期间的数据，所以对这一点就忍痛割爱，不做详细论述了。前文的这些结果可以归纳为如下内容（图 N）。

图 N 对问题的关键的汇总

	“快照”的视点	“趋势”的视点
大小 / 比例的角度	• 车型 B 与车型 C、车型 D 的平均客户忠诚度没有太大差异 • 只有车型 A 的销售数量比例较小	车型 B 与车型 C 的客户忠诚度自一年前起出现下降
波动的角度	车型 D 的波动较大	（非此次检验对象）

接下来，作为问题的关键进一步深入挖掘的对象，可以锁定为车型 B 与车型 C。一个重要的事实是，车型 B 与车型 C 的结果并不是通过偶然的调查碰巧得出的。我们从逻辑上探索最根本原因的同时，通过全方位的检验，将这 2 个车型锁定为问题的关键。面对“为什么最后锁定这 2 个车型”的质疑，能够提供确凿的“根据

（论据）”，这一点可以确保整个问题解决过程具有可靠性和逻辑性。

分析到这一步，接下来要针对“为什么车型 B 和车型 C 的客户忠诚度会出现下降”，进行原因分析。

第 3 章
采用交叉视点，锁定“原因”

前面介绍了如何用平均值和标准差等指标来“计算每月平均销售额”“比较每家店铺来店人数的波动”等。此外还介绍了如何在此基础上从某个维度对大小、比例或波动等进行比较，划定对象数据的范围，锁定问题的关键。[书籍分 享V 信zmxsh998]

那么，从这些指标可以获得哪些信息呢？不过是现有数据范围内的结果或者现状。进一步分解销售额的构成要素，对每种商品、每个店铺、不同月份等进行深入研究（分解），能够更进一步锁定具体的问题的关键。不过其结果仍然不过是“更为细致的现状”。

在实际工作中，我遇到的很多情况是：虽然通过数据分析细致入微地了解了现状，但到锁定原因的阶段，却又用突兀的主观见解来代替客观事实。“这一步也能展现出客观根据就完美了”，像这种令人深感惋惜的例子屡见不鲜。

通过在日产工作的经历，我切身感受到，接下来的这一步能否用数据说话，会导致后面的工作方法截然不同。因为人们对其认可程度完全不同。尤其是如果在某个关键点突然变成定性的主观逻辑，原本连贯的“故事”就会戛然而止，整个分析也会变得苍白无力。

在解决实际问题的过程中，正确了解现状作为第一步确实非常重要，但绝大多数情况下，这一步的最终目的应该是锁定导致该结果的原因，为解决问题而采取必要的措施（行动）。然而很多运用数据的过

程，其实都没有超出了解现状这一步。那么为什么无法从了解现状的阶段继续走下去呢?

这是因为人们的思维被限制在销售额、来店人数等单个维度的范围之内。不摆脱限制，就无法跨出了解现状的范围。假设使用平均值、标准差来把握“XX 商品的销售可能存在问题”的现状，或者将其锁定为问题的关键，那么接下来就应该考虑“为什么 XX 商品的销售额会下降”。

再接下来，除了销售额以外，还必须关注有可能影响销售额的种种原因（驱动因素）。例如可能包括产品、客户服务、宣传促销、价格政策等。知道了这些因素中有哪些影响销售额、哪些并无影响，说不定就能够锁定原因。

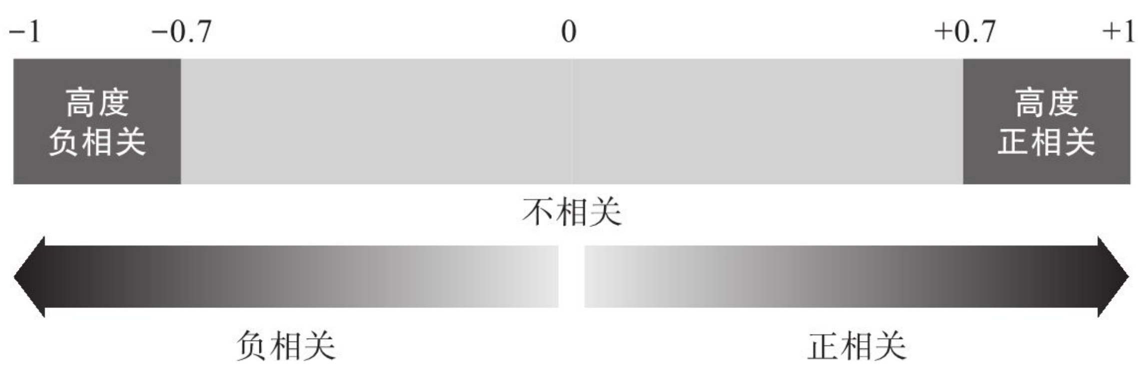
表示二者关系的“相关系数”

像这样，关注 2 种以上数据之间的关联程度，就可能获得仅从一种数据中绝对无法得到的信息。换句话说，就是将视野从一维的数据，扩大为关注 2 个维度，有意识地转换思路和视点。

为了把握 2 种数据之间的关联，首先需要确认它们是否密切相关。

“相关系数”可以表示相关程度，计算相关系数的方法叫作“相关分析”。相关系数的值介于-1 和+1 之间（图 3-1）。

图 3-1 相关系数



相关系数越接近 1，正相关的程度越高。也就是说，一方数据增加，另一方数据也会随之增加。二者完全成比例（如果一方增至 2 倍，另一方也随之变为 2 倍）时的相关系数最大，是 1。相关系数为 0，表示两个数据没有任何关联，互相独立。实际业务中使用的第一手数据，一般都不是 0 或者 1 所表示的完全不相关或者完全成比例相关，而是介于二者之间。

Excel 的小妙招

那么，相关系数是什么数值时，可以判断为“相关”呢？判断相关系数的大小并无一定之规，一般 0.7 以上可以视为“高度（正）相关”。0.7 并不是一个严格的标准，现实中根据不同的分析目的和所需准确度，有时 0.5 以上即可视为相关。

同样，当相关系数为负数时，两种数据则为“负相关”。负相关只是数据的变化方向不同，原理与正相关一样。也就是说，如果一方数据增加，另一方数据则会随之减少。

以下为相关分析的具体事例。例如，在根据预计来店人数来决定营销费用的情况下，来店人数与营销费用之间正相关的程度越高，营销费用的效果越好。投入的营销费用越多，来店人数就增加越多。但如果二者之间没有足够的正相关关系，营销费用就很有可能白白浪费。图 3-2 为运用 Excel 函数计算相关系数的例子。

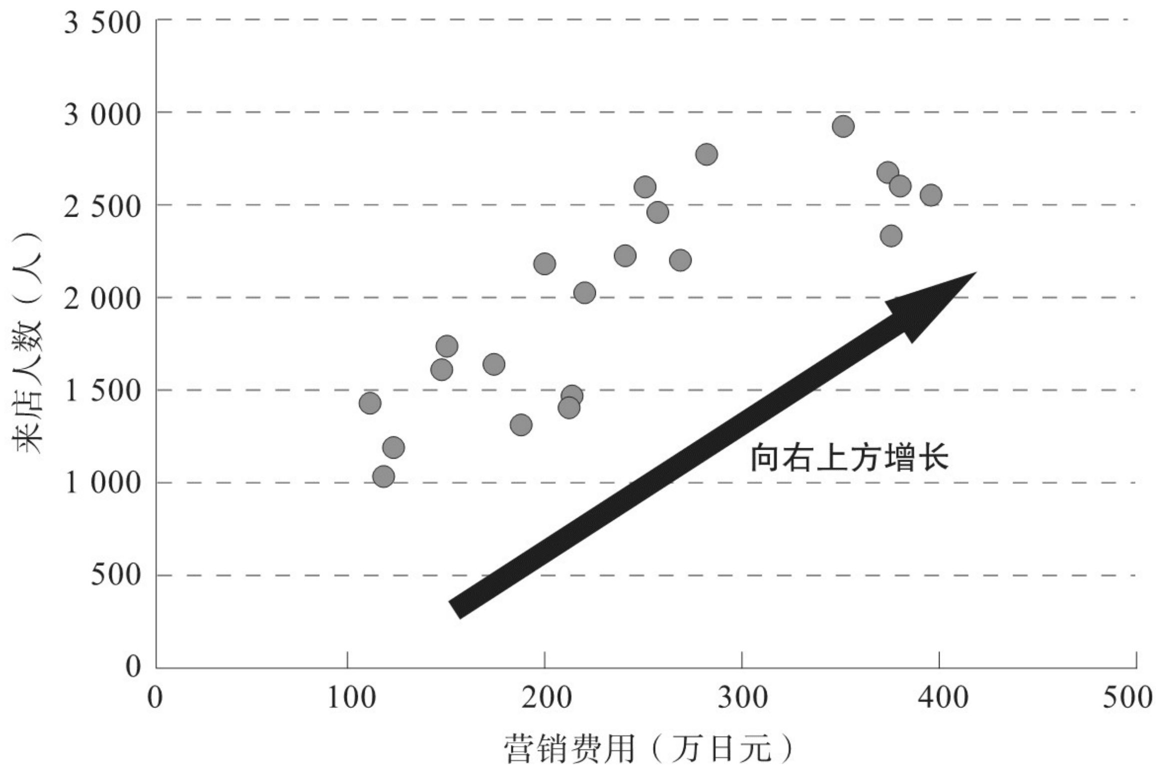
使用 CORREL 函数可以计算相关系数。在“=CORREL”后面的括号中指定两种数据的范围，并用逗号隔开，即可立即得出相关系数。在图 3-2 的例子中，相关系数为 0.84，可知二者高度相关。

图 3-2 来店人数与营销费用的变化

	A	B	C
1		来店人数（人）	营销费用（万日元）
2	2013 年 4 月	2 198	269
3	2013 年 5 月	1 035	118
4	2013 年 6 月	2 230	241
5	2013 年 7 月	1 614	148
6	2013 年 8 月	2 028	220
7	2013 年 9 月	2 606	379
8	2013 年 10 月	2 332	376
9	2013 年 11 月	2 550	396
10	2013 年 12 月	1 735	151
11	2014 年 1 月	2 589	251
12	2014 年 2 月	1 634	174
13	2014 年 3 月	1 425	111
14	2014 年 4 月	2 184	200
15	2014 年 5 月	2 669	374
16	2014 年 6 月	1 468	214
17	2014 年 7 月	2 767	282
18	2014 年 8 月	2 920	351
19	2014 年 9 月	2 458	257
20	2014 年 10 月	1 312	188
21	2014 年 11 月	1 413	212
22	2014 年 12 月	1 186	123
	相关系数	0.84	=CORREL(B2:B22,C2:C22)

作为参考，我们再看运用这些数据制成的散点图（图 3-3）。纵轴为来店人数，横轴为营销费用。由于具有 0.84 的高度相关，纵轴会横轴的增加而增加，从直观上也可以确认到向右上方上升的趋势。

图 3-3 营销费用与来店人数相关



如果两个数据不相关，散点图就会呈现出不规则分布。因此除了相关系数之外，用散点图从直观上展现两种数据之间的关系，有时也可以发挥重要的作用。尤其对下面 3 种情形来说，散点图会很有用。

①能够发现明显的离群值（出于某种原因，明显偏离其他数据的数据）。离群值可能会产生影响，导致整体的相关系数变低。如果能够合理去除离群值，那么其他数据的相关系数可能会有不同。

②相关系数能够体现两种数据之间的比例关系（线性关系），但并非所有数据之间都是比例关系，也可能是其他类型（曲线等）关系。这些其他类型的关系可以不依赖相关系数，从散点图中看到。

③借助散点图，对相关分析、相关系数一无所知的人也可以理解分析的结果。

并非所有情况都需要制作散点图。可以先通过 CORREL 函数计算出相关系数，然后再针对重要内容用散点图加以确认。

锁定原因也需要“假设”

第 1 章中曾经提过，着手处理数据之前，应该首先确认目的或问题，在此基础上提出假设，这一点非常重要。在分析问题原因时，假设也同样有效。与原因有关的假设叫作 WHY 型假设。

例如，在刚才的例子中，目的或者问题是“增加来店人数”。为此首先需要找出影响来店人数这个被解释变量的原因。一般来说，在自己所从事的行业或业务中，找出可能影响最终结果的原因并非难事。

但需要提防一些陷阱，防止受前例束缚的主观臆断，或者将视野限定在平时常见的数据范围之内。

克服常识的限制，不因为“这是平时一直关注的数据”“前辈、上司都这么说”而放弃思考，而是客观地进行探索，这也是从事数据分析的条件之一。此外，还有 2 点需要注意。

（1）寻找接近结果的原因

例如，针对提高销售额的目的，将“营销费用”作为原因的话，销售额和营销费用需要通过“来店人数”才能具有关联。三者之间是“销售额—来店人数—营销费用”的链条结构。

但实际上不只有来店人数会影响到销售额，而营销费用与销售额的直接关联就更为薄弱。所以这两个数据之间的相关关系就不太容易把握准确。因此，最好先分析类似来店人数与营销费用等关系更为紧密的数据之间的关联，在关联更密切的数据之间考察相关关系。

（2）选择能够采取对策的原因

即使可以从理论上找到很多原因，并断定其与要实现的指标具有高度相关，但如果都是自己无法控制的外界原因，那么这个分析结果就派不上用场。如果分析的目的就是研究倒也无妨，但如果是为了在实际工作中实现目标或者解决问题而做数据分析，这样就会毫无意义。建立假设时就要设想“假设得到验证，就能采取哪些对策”，这也是考验数据分析者能力的一个方面。

相关系数的四大优势

如果你过去只会从平均值或图表中探索数据含义，那么今后在锁定问题原因时可以使用相关分析这个得力工具。这是帮助数据分析在实际业务中向前迈出一大步的最佳方法之一。因为相关分析具有以下四大优势。

（1）能够立即得出答案

正如前文介绍的，运用 Excel 的 CORREL 函数，立刻就能得出相关系数。对每天忙于日常业务，需要在短时间内找到答案的一线工作人员来说，相关系数可以节约时间的特点极其重要。在短时间内得到答案，就意味着可以在有限时间内多次尝试。

我们不一定从一开始就能得到预想的结果。所以进行数据分析时，能够在短时间内检验多个假设，这个优势不容忽视。

（2）简单易懂，更容易得到对方理解

相关分析的结果（相关系数）介于-1 至+1 之间，对任何人来说都很易于理解。而且无须向对方详细解释“相关”的含义（其他分析方法的名称大多令人费解）及分析结果的意义。实际工作中，数据分析的结果需要得到理解和认可，因此是否便于沟通也很重要。在这一点上，相关分析可以说是完全合格的。

（3）能够分析单位不同的数据

在工作中，输入的信息与输出的信息不一定都能用同样的单位表示。代表性例子是，发送网页宣传单的次数与来店人数之间的关系。输入信息是发送网页宣传单的次数，单位是“次”，而输出信息是来店人数，单位是“人”。两种数据单位不同，无法进行四则运算，或者单纯用数值进行比较，但做相关分析就完全没有问题。也就是说，相关分析的方法通用性很高，不受单位限制，可广泛用于现实业务。

（4）为回归分析等进一步分析做铺垫

将数据的相关关系写成公式，就可以通过“回归分析”（第 4 章）做出更为具体的预测或制定计划方案等。也就是说，相关分析还可以为下一个分析阶段提供线索。

相关分析可以单独运用于实际业务，也可以为其他深入分析做准备。这样可以避免只用单独的一种方法结束分析，而是用其他方法做补充，或者组合使用，从而通过多个分析形成脉络，描绘出解决问题的完整故事。正如我多次强调的，贯穿始终的故事可以增强分析的论据，显著提高对方的理解和接受程度。

我在日产工作时，曾在很多项目中运用过相关分析。对那些全面掌握商业管理领域各项技能的高管，不用解释“什么是相关”，相关分析可以被所有高层管理者积极地接纳。即使有人不了解相关分析，也可以使用散点图直观展现两个数据之间的关系，获得相同的效果。散点图的直观效果与相关系数的定量分析相辅相成，在很多场合都曾发挥出卓越效果。

除了需要高层管理者进行决策的场合，在销售部门的日常业务中，相关分析也能够发挥威力。一般来说，汽车行业拥有丰富的销售数据。我刚入职时，也会收到各种途径汇报上来的销售业绩，但这些数据只不过是从各种角度展示的业绩而已。半路入行的我最感兴趣的是，这些数据与结果之间具有怎样的联系、会受到怎样的影响等问题。因此我使用各种数据，逐一考察它们与结果类数据之间的相关关系。

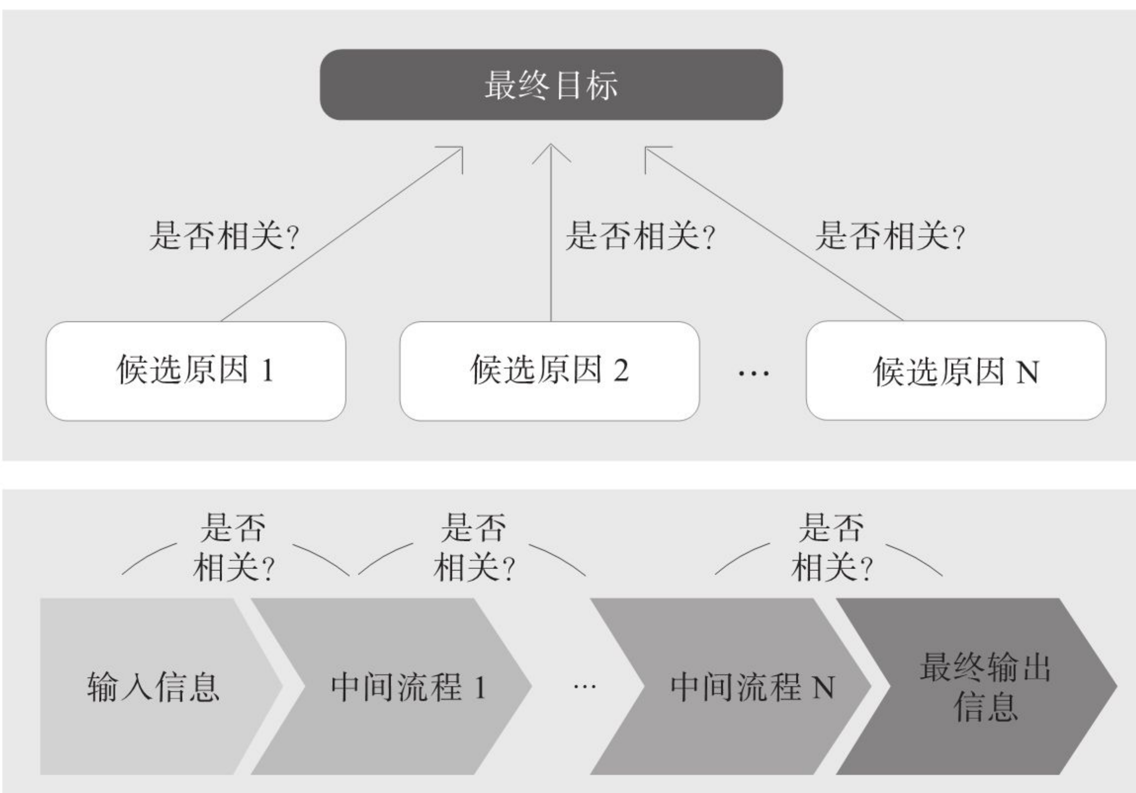
这么做的结果，是一些之前“只是结果”的汇报可以发挥更大的作用，使我洞悉影响结果的原因。我所属的部门负责区域销售，过去没有

将数据组合在一起看过，而这种做法作为一个新突破口或新视点，成为有效利用数据的第一步。

找到相关分析的着眼点

那么，应该如何将相关分析用于眼前的问题呢？有一个降低难度的方法，就是记住常见的应用模式。图 3-4 列举了在解决问题的实际工作中应用相关分析的两种模式。

图 3-4 在实际业务中运用相关分析的着眼点



第 1 种模式通过相关分析，找出对最终目标具有密切影响的原因。例如，针对“销售额”这个最终目标，（通过假设）找出“顾客满意

度”“降价”“产品魅力”等可能影响该目标的因素，观察这些因素与其相关程度的高低，可以得知按动哪个按钮（原因），会对销售额产生剧烈影响。这样就可以将资源转移到与目标关系更为直接的活动中。反之，对那些本以为有助于增加销售额而一直进行的活动，如果未能发现它们与目标的相关关系，结论可能就是需要重新考虑是否继续这些活动。

日产公司的全球总部也会面向各地区开展各项策划或活动。但寄予厚望而策划的活动，有时却未能如愿，或者最初虽然有效，但几年后就变得收效甚微。负责与当地沟通的部门，常能听到来自一线的类似反馈。

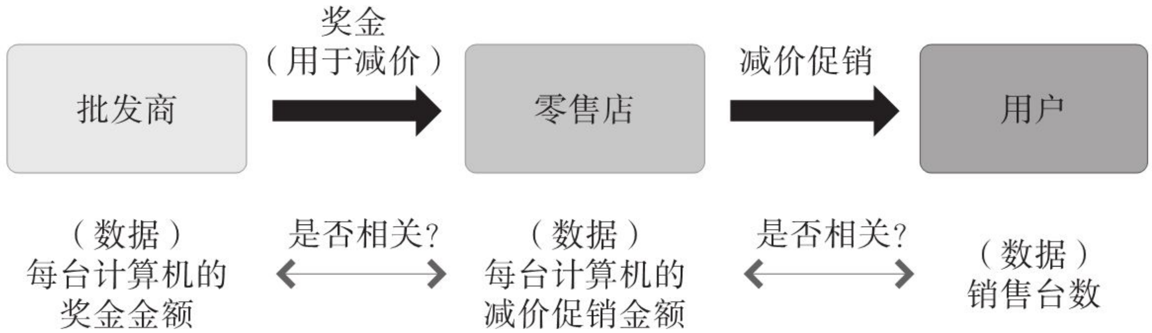
实际上，对正在开展的活动，很难对其成果展开讨论，或者在讨论之后终止活动。有时开展活动这件事本身更容易被视作成果，无法客观证明“已经没有效果”，是很难决定终止活动的。

这种情况下，新的项目接二连三地分派到当地，却无法确定哪些项目可以不必再继续下去，当地可能就会陷入项目不断增加的困境。为了解决这个问题，我们曾经多次尝试用相关分析来检验效果。这样一来，不是由某个个人做决定，而是通过“数据”“客观地”展现效果的有无，从而反映出真实情况。用数据区分有效的项目和无效的项目，可以对有限的资源实现最优化配置。

第 2 种模式是在某个业务流程中找到瓶颈。业务流程中包含金钱、信息等的流动，理想的状态是所有要素都畅通无阻地抵达最终阶段（输出）。但如果某个环节出现了停滞，预想的输出就无法实现。让我们来看一个具体事例。

假设某计算机批发商向各零售店支付奖金，以便它们在各自店铺进行减价促销。零售店将奖金用来减价，从而增加销售。其具体流程如图 3-5 所示。

图 3-5 在销售过程中应用相关分析



面对“计算机销售未取得预想业绩”的问题，需要在实现“销售台数”这一最终输出的流程中，找出问题（瓶颈）所在。找不到问题，就无法采取适当的措施。按照设想，本来应该是以下情形：

•“每台计算机的奖金金额”越高，零售店“每台计算机的减价促销金额”越高

•“每台计算机的减价促销金额”越高，购买计算机的顾客越多

其中任何一项存在问题，在“每台计算机的奖金金额”与“每台计算机的减价促销金额”之间，或者“每台计算机的减价促销金额”与“销售台数”之间就会看不到相关关系。

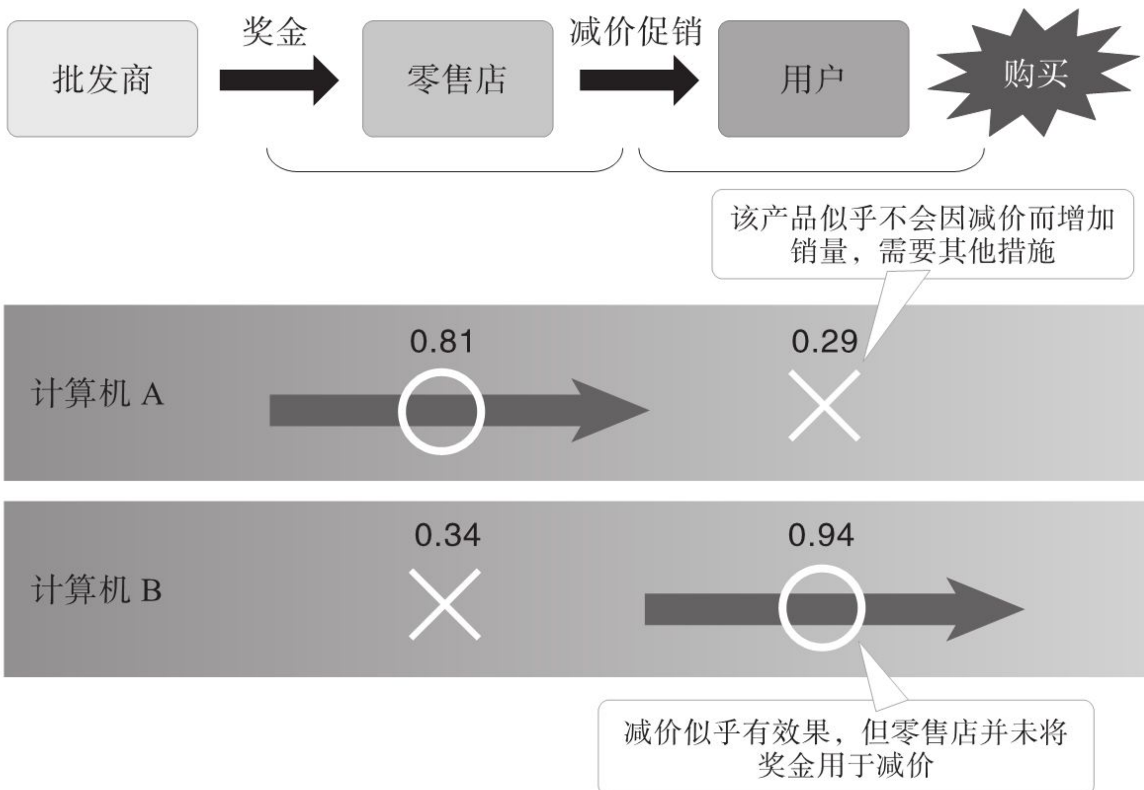
奖金被用到哪里了

进行相关分析之后，各数据之间可以得出如图 3-6 所示的相关系数。

计算机 A 的情况是，奖金被用于减价促销（相关系数：0.81），但减价并未体现为销售台数的增加（相关系数：0.29）。看来计算机 A 并不能依靠减价的方法来增加销量。这样，就找到了店铺减价流程中的一个瓶颈。计算机 A 需要立即改为采取减价以外的其他措施。

计算机 B 的情况是，似乎只要减价就可以期待销售台数增加（相关系数：0.94），但零售店并未将奖金用于减价促销（相关系数：0.34）。这样，就发现了零售店在减价流程中造成的瓶颈，也许是零售店对一线员工管理不利造成的，也许是零售店故意将奖金揣进了自己的口袋。此时需要采取的措施是对零售店加强指导，或者将销售渠道改为其他零售店。

图 3-6 相关分析的结果



像这样，关注流程之间传递的数据，分析它们之间的相关关系，有时可以发现需要解决的问题。在实际工作中应用相关分析，一开始可能比较难，但如果我们随时意识到图 3-4 所示的 2 种模式，将其套用到各种情形，就能通过反复尝试逐渐培养出这种视点和直觉。

相关分析探索的是两种数据之间的相关系数，但实际业务中，需要考察的对象可能不只两种。例如下面的情形，假设这是过去 15 周内各种商品销售数量的数据（图 3-7）。

图 3-7 过去 15 周内各种商品销售数量

周	冷冻食品	酒	点心	副食	面包
#1	458	450	330	228	432
#2	162	283	317	166	478
#3	418	325	269	245	104
#4	502	439	169	217	344
#5	390	350	274	253	415
#6	171	154	318	199	194
#7	560	463	135	411	256
#8	193	278	217	157	126
#9	222	183	153	364	179
#10	391	422	210	446	151
#11	234	293	239	193	277
#12	197	255	444	306	481
#13	416	195	397	116	489
#14	393	455	175	426	109
#15	368	465	346	398	478

用 Excel 加载项进行批量分析

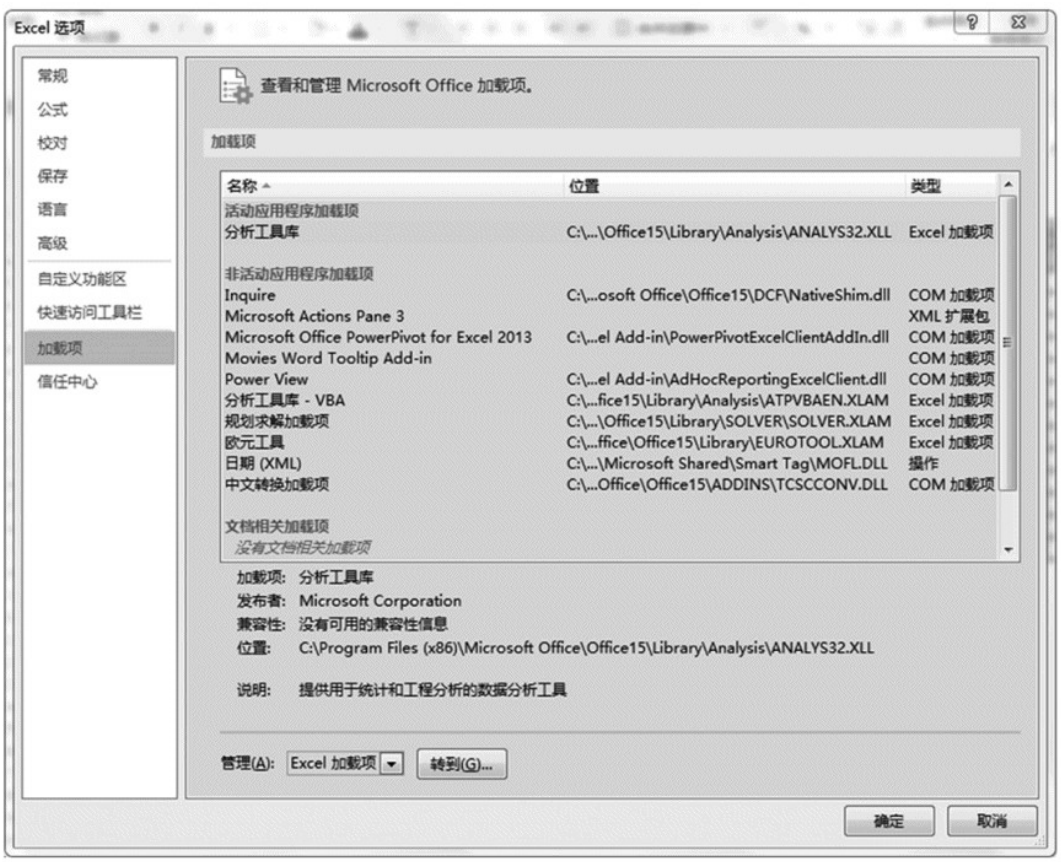
图 3-7 的样本量在实际工作中其实还不够充分（一般认为最少需要 30 个样本），不过我们就暂且用它来粗略地考察一下各商品销售数量是否相关。然后可以根据其结果提出假设，在卖场找到有利于顾客依次购买的摆放方式。从冷冻食品到面包，5 种商品两两组合，可以得出 10

种模式。对这种规模的数据，可以用 Excel 的 CORREL 函数逐一计算出相关系数。

但如果商品种类变为 10 种，就会产生 45 种组合模式。这时再用 Excel 的 CORREL 函数计算相关系数会非常吃力。这种情况下，可以使用 Excel 标准配置的“加载项”功能。这里以 Excel 2013 为例进行说明。其他版本可能略有不同。

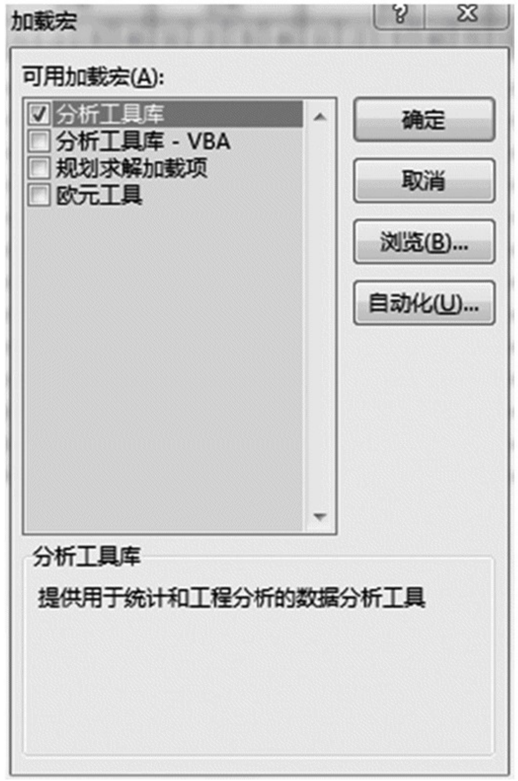
首先在“文件”菜单中单击“选项”。在接下来出现的页面点击“加载项”，并从右侧列表中选择“分析工具库”（图 3-8）。

图 3-8 选择加载项



点击页面下部的“转到”按键，在接下来的画面中勾选“分析工具”，点击“确定”。这样就在 Excel 中加载了分析工具。加载成功以后，同一台计算机以后无须再次加载，非常方便。

图 3-9 勾选分析工具



加载完成后，“数据”标签右侧会显示“数据分析”。点击“数据分析”，在图 3-10 所示页面选择“相关系数”。点击“确定”后，将数据所在范围指定为“输入区域”，在“标志位于第一行”前划✓，点击“确定”完成操作。于是前文的例子就可以得到图 3-11 所示结果。这里将小数位数设定为两位。

图 3-10 选择“相关系数”

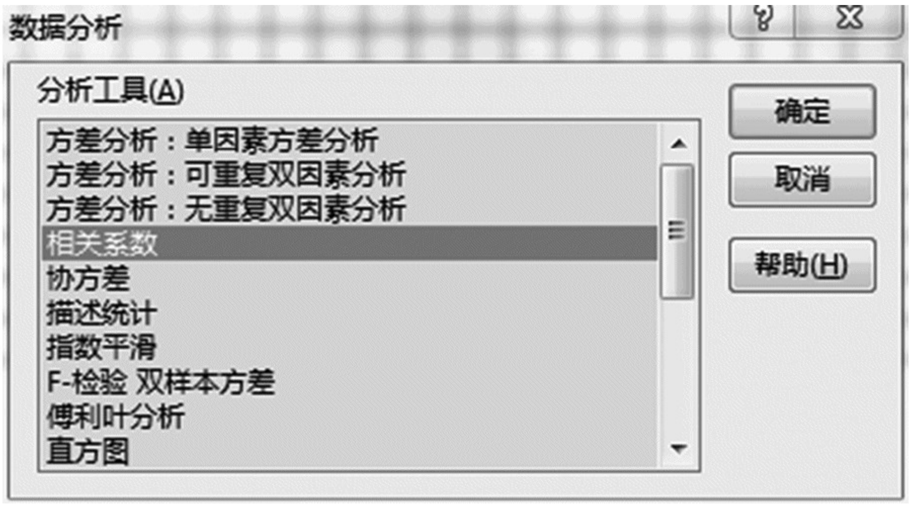


图 3-11 5 种食品的相关系数矩阵（分析结果）

	冷冻食品	酒	点心	副食	面包
冷冻食品	1				
酒	0.72	1			
点心	0.31	0.34	1		
副食	0.32	0.54	0.40	1	
面包	0.03	0.01	0.69	0.30	1

图 3-11 显示了 5 种数据之间所有组合的相关系数。从中可以发现，有 3 种组合，即冷冻食品与酒、酒与副食、点心与面包的相关系数超过 0.5。可能是最近独自一人吃晚餐的顾客比以前增加了，酒与冷冻食品或副食的销售情况比较相似。点心与面包的组合可能是面向儿童顾客的。进一步调查星期几、哪些商品或者哪些时间段的销售情况更好等信息，也许可以得到更深入的结果。

这样一来，很容易就可以找到下一步分析的线索。数据种类较多的相关分析中，Excel 加载项是一个强大的工具。对数据分析感兴趣的读者，不妨立刻启动加载项尝试一下。

利用矩阵排列优先顺序

相关系数不仅能单独使用，还可与其他指标组合起来进一步应用。例如，图 3-12 的例子是温泉旅馆的顾客问卷调查结果。除了综合满意度之外，该问卷调查还包括顾客对其中 5 个因素的评分。

图 3-12 温泉旅馆顾客满意度问卷调查结果

顾客编号	综合满意度	料理	价格	洗浴	接送	服务态度
#1	62	68	77	55	97	53
#2	36	65	94	50	22	22
#3	93	64	48	67	64	86
#4	41	67	58	32	18	87
#5	76	91	100	58	75	56
#6	34	51	41	54	29	35
…	…	…	…	…	…	…
#26	78	88	90	48	75	55
#27	64	88	77	27	47	50
#28	53	69	55	22	68	45
#29	72	95	68	62	92	99
#30	63	88	61	39	42	52

“综合满意度”决定了顾客是否会再次光临，因此需要判断在哪个方面加大力度才能提高综合满意度。旅馆的资源（经费、时间、人员等）有限，需要对 5 个因素进行比较，决定其优先顺序。为此，我们考察了各因素与综合满意度的相关系数（图 3-13）。

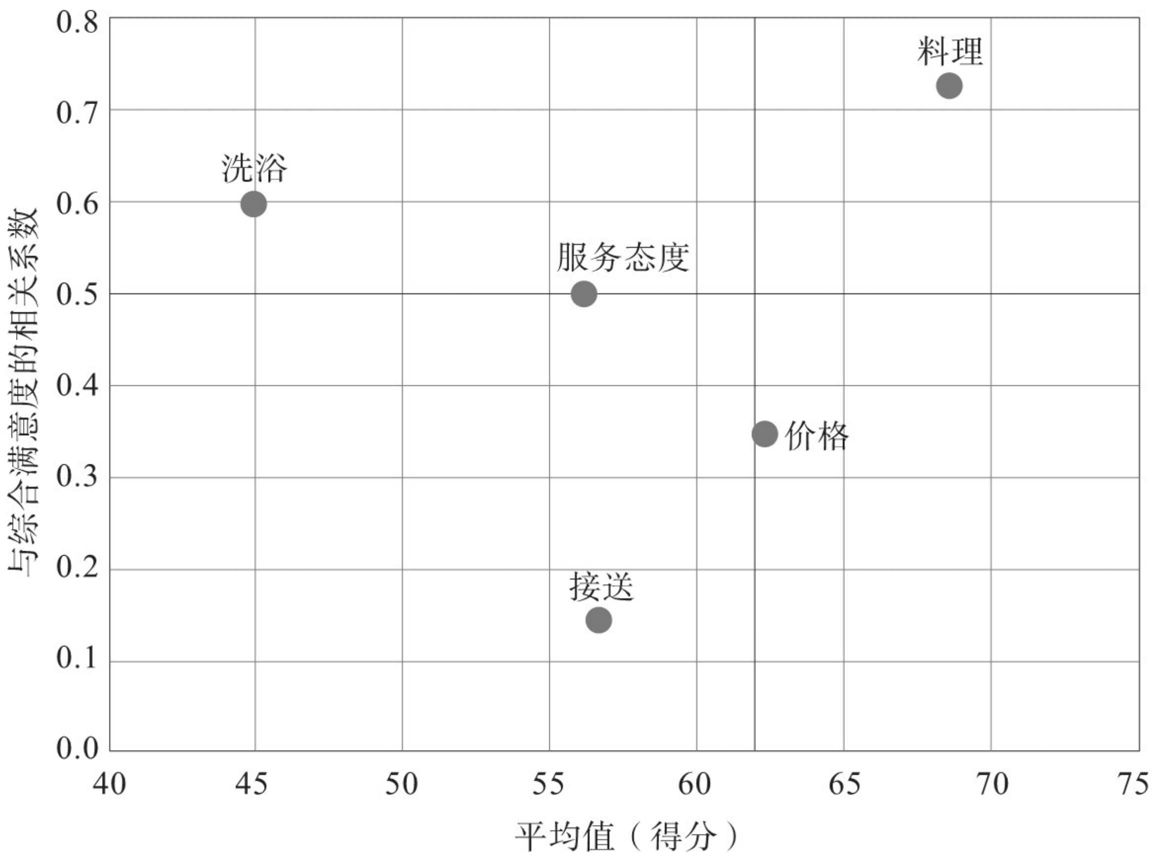
图 3-13 各因素与综合满意度的相关系数

	综合满意度	料理	价格	洗浴	接送	服务态度
与综合满意度的相关系数		0.73	0.35	0.60	0.14	0.50
平均得分	61.6	68.7	62.2	44.8	56.7	56.2

同时还计算了各个因素的平均得分。仅凭这些仍然难以做出判断，所以又以相关系数为纵轴，以平均值为横轴，做成图 3-14 所示的矩阵。

为了明确评价标准，我们将综合满意度的平均值（61.6 分）和相关系数 0.5 做了突出显示。与综合满意度具有相关关系（以 0.5 为标准）的，包括“料理”“洗浴”和“服务态度”。“料理”与综合满意度的相关程度最高，不过“料理”的得分已经达到了较高水平，远远高于平均值，因此与其他方面相比，可改进的余地比较小。当然，如果“料理”得分降低也会产生严重的后果，所以必须要维持现状。

图 3-14 顾客满意度矩阵



另一方面，“洗浴”与综合满意度有相关关系，但平均得分较低。因此“洗浴”方面还有较大的改进余地，而且改进越多就越能提高综合满意度。同理，继“洗浴”之后需要强化的是“服务态度”。

需要注意的是，这里将与“综合满意度”的关系的密切程度作为一个维度，将“得分”数值作为另一个维度，这样可以把性质迥异的 2 个维度结合起来，互相补充对方所欠缺的信息。

也就是说，只靠相关系数体现的关系密切程度，无法看出“现在的得分是高还是低”。另一方面，仅凭“得分的高低”，也无法了解它对综合满意度有多大贡献（恐怕很多人都有过仅凭得分高低做判断的经历吧）。将 2 个维度组合起来，可以补充缺失的信息，能够更深入、更清晰地体现出整体情况。

像这样，需要客观地决定优先顺序时，可以通过 2 个维度展现其相对定位。这样不仅可以使分析更具深度，也更便于与其他人共享直观化

的信息，在组织内形成共识。请想象一下对每个维度单独展开讨论的情形，与只用平均值单纯比较相比，运用相关系数得出的结果更易于展开讨论。

来自不同部门或具有不同国籍等多样化背景参与者之间，很难实现相互理解并达成共识。不过有一些方法确实可以让所有人都更容易接受。“2 个维度”的直观方法就是其中之一。它最大的优势就是，既可以让人从主观上理解，同时又能够提供客观依据。

我曾经多次使用过这个武器。例如有时需要在短短 5 分钟的有限时间内，同时展示结论和根据，征得所有高层管理者的一致同意，那么这个方法就可以发挥绝佳效果。

不要随便编故事

相关分析既有效又简单，稍做尝试即可得出结论，非常适合实际应用。但另一方面，如果理解或使用方法不当，相关分析也有可能得出错误的结论。如果能在分析过程中发现这些错误倒也无妨，但这一点往往很难做到。实际上，有很多情况必须依靠分析者本人的细心、知识或经验才能发现。没有一种方法可以保证“这样做绝对没问题”。

因此，我们在学习分析方法的同时，还需要了解它的注意事项。随时有意识地进行检查，可以大大提高数据分析的质量。那么，都有哪些陷阱呢？以下介绍 4 种常见的情况。

陷阱 1　因果关系

得知“具有密切关系”之后，人们常会下意识地把某些故事套在上面。其中最容易接受的故事就是因果关系。将信息碎片随意拼接起来，

编造出若有其事的故事（因果关系），然后再用它来“说服”自己，这种情况在日常生活中十分常见。进行客观分析时，尤其需要注意这一点。

例如，“媒体曝光度”与“咨询件数”之间似乎具有高度相关。那么是否就可以得出结论，认为“进一步增加媒体曝光，就能增加销售额”呢？答案是“也许是，也许不是”（图 3-15）。

也许是由于“咨询件数”的增加，引起同行业的关注，才导致了“媒体曝光度”增加这一结果。

图 3-15 媒体曝光度与咨询件数之间是否具有直接关联？



相关关系并不一定就是因果关系，这一点非常重要。是否给相关关系套上因果关系的故事，是分析者个人判断决定的。所以要清楚，这并不是分析结果所展现的内容。根据同样的分析结果，却有可能得出与原本的因果关系截然相反的解释。

例如，二氧化碳浓度与气温变化之间确实存在相关关系。人们一般认为，“因为二氧化碳增加，导致地球变暖”，但其实也有论文提出截然相反的观点，认为是因为气温升高，使海水中的二氧化碳被排到大气中，导致二氧化碳浓度上升。在这个例子中，对于哪个是原因，哪个是结果，人们做出了不同的解释。仅靠相关分析，无法明确其因果关系。

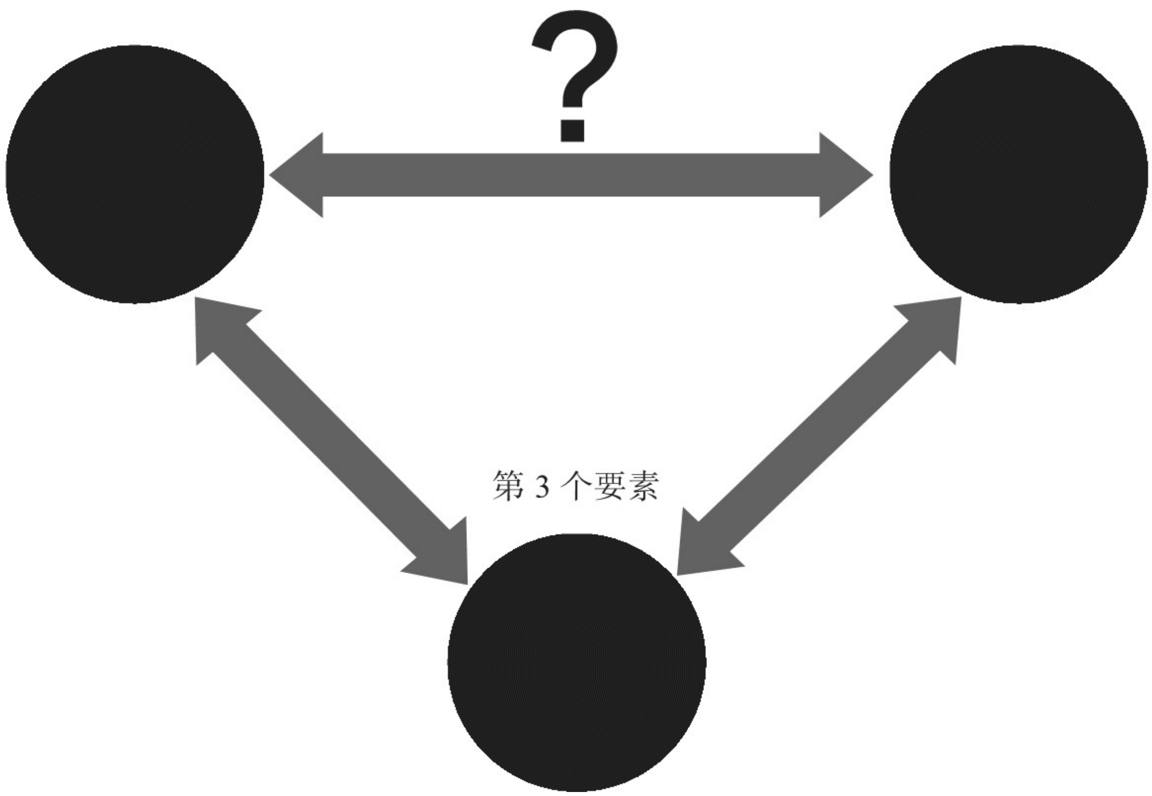
作为检验因果关系的方法之一，可以考察 2 种数据发生变化的时期（顺序）。因果关系应该是原因变化在先，结果随后发生变化。

陷井 2　疑似相关

任何情形都可以通过计算得出相关分析的结果（相关系数），但这个结果未必都是由“直接”相关关系导致的。尤其是在复杂的商务世界中，只用一对一的直接关系就能解释的情况其实很少。如果仅凭计算结果创造故事，很可能会得出有悖常识的结论。

例如，假设“顾客满意度”与“销售额”高度相关。那么是否就能直接得出“改善服务，提高顾客满意度，就能进一步提高销售额”的结论呢？对此，答案依然是“也许是，也许不是”。仅凭相关分析的结果无法得出真正答案。如图 3-16 所示，这种情况要求分析者考虑是否存在“第 3 个要素”。

图 3-16 顾客满意度与销售额之间可能存在“第 3 个要素”



在这个例子中，如果还存在分析中没有体现出来的“减价”，就不能否定减价导致销售额和顾客满意度同时提高的可能性。顾客满意度与销售额同时增加，数字上得到的结果显示两者高度相关，但我们却不能就

此认定二者有直接关系（例如因果关系）。在这种情况下不能把思考停留在眼前的数据上，还要随时提出多种假设，分别考察它们之间的相关关系，才能从某种程度上降低风险。

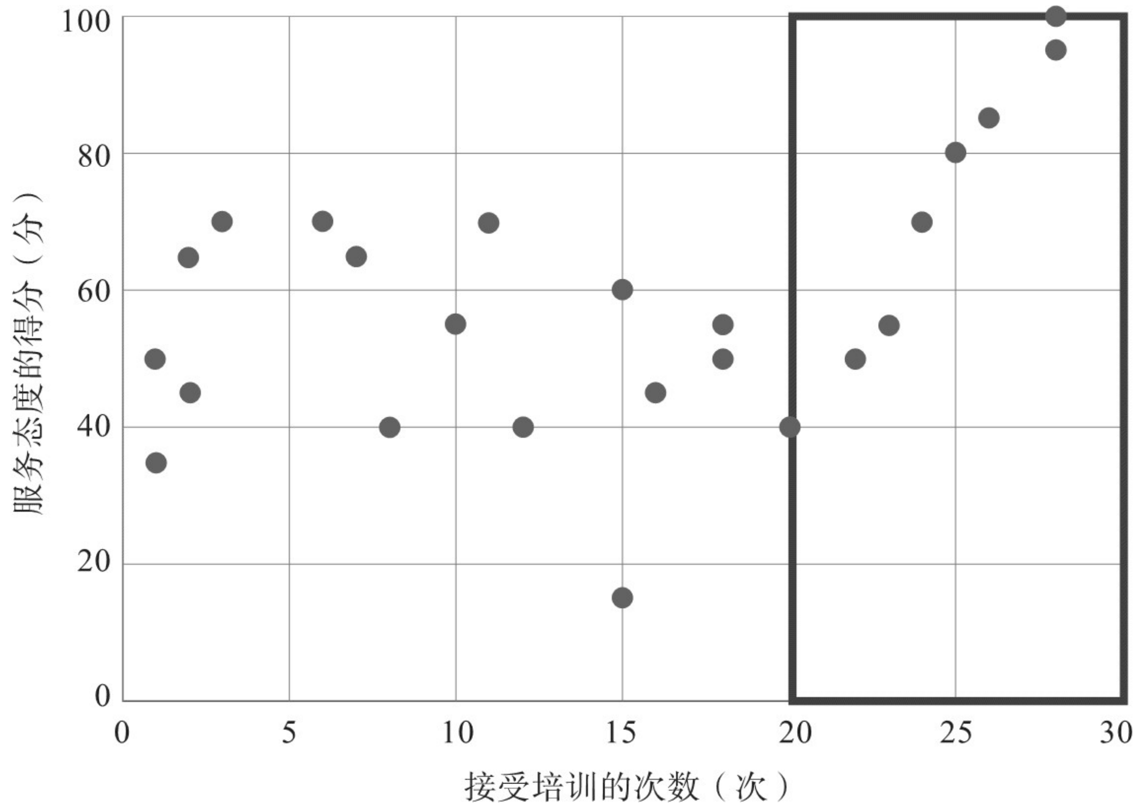
如何解释分析结果，需要分析者亲自判断。扩大解释的选项范围，可以避免结论受到狭隘思路的限制。

陷井 3　数据的范围

所有的数据分析都有一个共同点，就是分析所用的数据范围不同，会对结果带来很大差异。图 3-17 显示了某保险产品销售专柜员工“接受培训的次数”与“服务态度在顾客问卷调查中的得分（服务态度得分）”的关系。如果一股脑地计算所有数据的相关系数，可以得到 0.40 这样一个并不算高的数值。

不过我们并不能据此得出“接受培训次数的多少对顾客满意度没有贡献，因此没有意义”的结论。因为如果假设“只有达到一定次数之后，培训效果才会显现”，那么就有可能在散点图上发现前后出现变化的点。图 3-17 的例子很明显，以接受培训 20 次左右为界，服务态度的得分趋势开始发生变化。也就是说，需要对分界点的前后加以区分，重新进行相关分析。

图 3-17 培训的效果检验



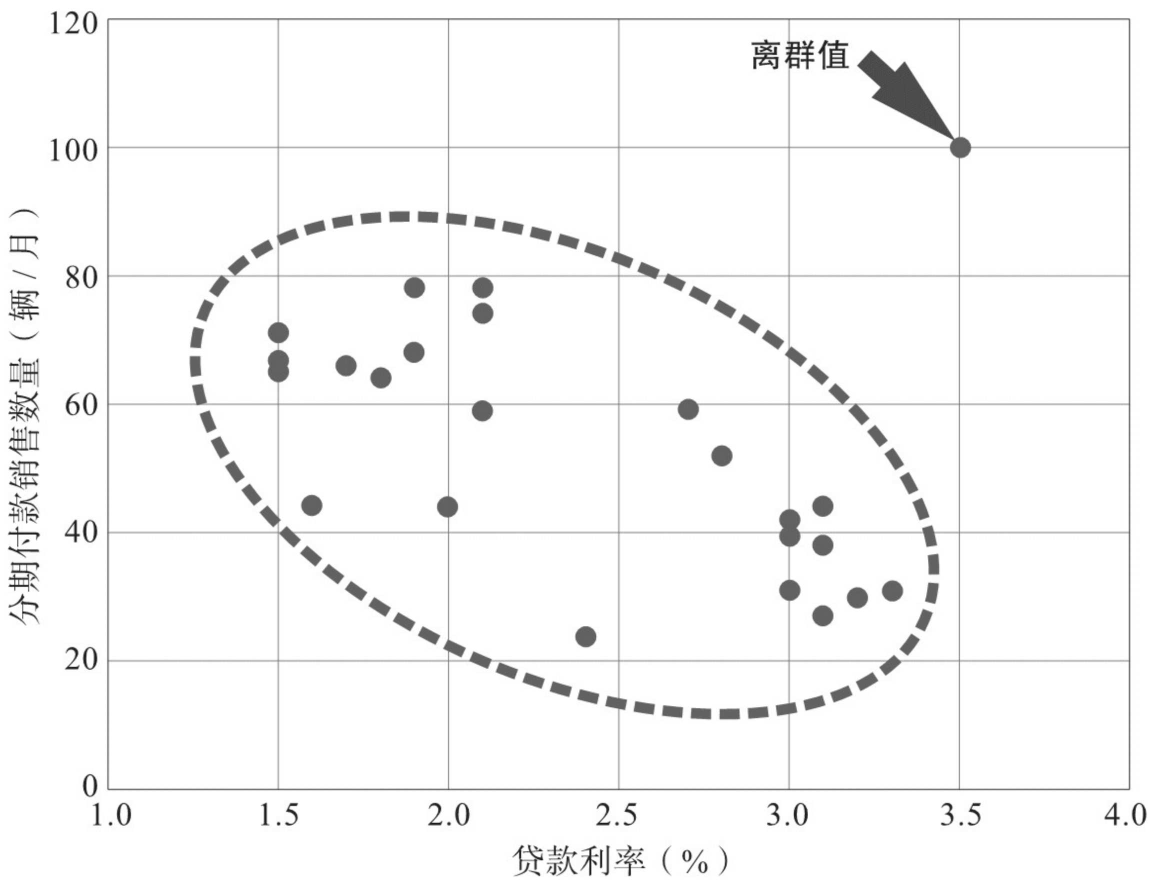
单独计算接受 20 次培训以后的相关系数,可以得到接近 1 的高度相关的数值。这个“选择分析范围”的工作也要由分析者来做。分析者能否针对数据范围提出适当的假设，有可能会导致截然相反的结论。

除了最初就知道应该着眼于何处的情况，一般情况下，可以用散点图将数据的相关关系直观地展现出来，然后再逐一探索应该从何处着眼，这也是一个关键。

陷井 4　离群值

还有一个与其他分析方法共通的注意事项，即对“离群值”的处理。离群值指由于某种原因，与其他数据差距比较大的数据。分析对象中是否包含离群值，会使相关系数产生很大差异。图 3-18 是从 25 家经销商收集的汽车分期付款销售数量。从整体上可以看出，随着贷款利率（横轴）的升高，分期付款销售数量（纵轴）呈下降趋势。

图 3-18 分期付款销售业绩



在右上方，有一家店铺并不符合这个倾向。它就是离群值。连同这个离群值一起计算，得到的相关系数为-0.44，并不太高。但如果剔除离群值，相关系数就会猛增为-0.74。

那么是否应该把离群值全都剔除在外呢？答案是“原则上，没有明确理由是不可以的”。

如果可以随意删除数据，就有可能出现分析者操纵分析结果，导致分析丧失客观性和可信度的情况。所以，发现离群值时，首先要调查这个数据为什么会出现偏离。在此基础上，如果能找到合适的理由，则可以将其从对象中剔除，再进行分析。这个事例中，由于只有右上方的经销商实施了极为优惠的首付政策，所以分期付款销售数量的增长与贷款利率无关。

收集此类无法从数据中获知的信息，并进行判断，也是分析者的重要工作之一。像图 3-18 一样，用散点图加以直观展现，会更容易发现离群值。

综上所述，与其他方法相比，相关分析有很多需要分析者多加留意、开动脑筋的地方，这与它能够用于各种问题和情况的较高通用性互为表里。虽然分析者的技术和直觉可能永远也达不到满分的水平，但随着分析经验的增加，其能力自然也会不断提高。

「解决问题的故事 3」

第三步：建立 WHY 型假设，关注影响客户忠诚度的要素

锁定问题关键并进行检验之后，确定车型 B 与车型 C 的客户忠诚度自一年前开始下降，导致很多顾客再次购买时转为选择其他公司产品，影响了销售数量和销售额。

那么为什么会出现这种情况呢？可以通过关注客户忠诚度与其他数据之间的相关关系来锁定原因。

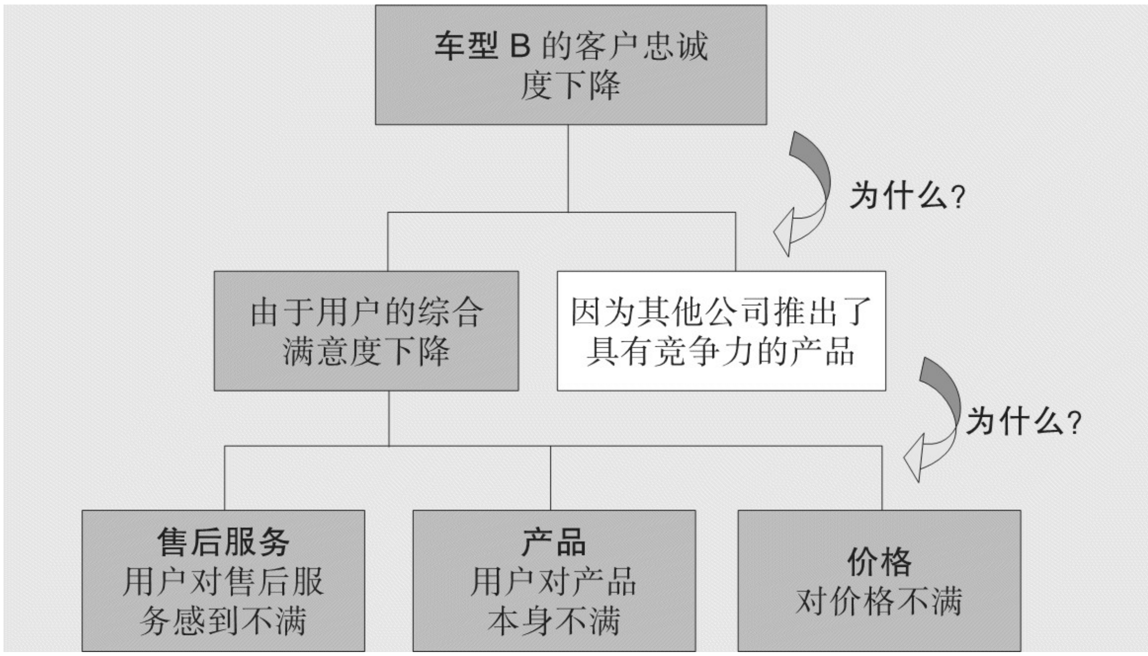
虽然也可以对客户忠诚度和可能相关的现有数据逐一进行相关分析并建立假设，但为了保证思路不受数据或随机因素的影响，我们首先从理论上进行思考。

要寻找原因，可以构建“WHY 型假设”。WHY 型假设是将问题放在最顶层，然后列出能够回答“为什么”的所有可能答案。接下来，再针对这些答案，同样重复“为什么”的提问，进一步深入挖掘。

图 A 从“用户再次购买时为什么选择其他公司产品”的疑问开始，最终将问题归纳为“售后服务”“产品”和“价格”3 个关键词。总

结关键词，可以便于检查是否有遗漏或缺失。一句话不太容易在大脑中留下印象，而只列出关键词，则比较容易发现遗漏或者重复。

图 A 车型 B 的客户忠诚度为什么会下降（WHY 型假设）



至于需要深入挖掘到何种程度，我认为只要最终能够锁定原因，并设想出具体措施即可。这个事例中，有一个原因选项是“因为其他公司推出了具有竞争力的产品”。对这个原因，无论如何深入挖掘，解决方法也只能是“开发更有竞争力的产品”等中长期措施，不能成为尽快增加销售的方法。因此，可以暂且将其优先顺序推后。

也有可能这才是根本原因，所以我们不是无计可施就忽视它的存在，而是因为目前需要优先调查能在短期内采取对策的原因。与那些需要严密调查理论上的所有原因，写成报告的情况不同，商务人士需要根据目的、制约条件和实际情况，采取灵活的措施。

那么，先来看综合满意度（月份平均）与客户忠诚度是否相关。因为没有区分不同车型的满意度数据，只有包括所有车型的综合满意度，所以需要计算它与各车型客户忠诚度之间的相关系数。

如图 B 所示，整体客户忠诚度与综合满意度之间的相关系数为 0.64，由此可知一般来说（不区分车型），两者之间存在相关关系。再看不同车型客户忠诚度与综合满意度的相关，车型 B 和车型 C 与综合满意度的相关系数分别为 0.75、0.69，数值较高，可以确认为相关。

此外，没有被列为问题关键的车型 A 和车型 D 与综合满意度不相关，决定顾客再次购买时如何选择的，可能是竞争对手产品等其他原因。

只看综合满意度，并不能决定"应该采取哪些措施"。这样的话仍然无法对实际业务产生意义，所以接下来还要再次应用相关分析来探讨"售后服务""产品""价格"与综合满意度之间的相关程度。每一种车型都有数据，所以就按照不同车型分别来看（图 C、图 D）。

图 B 不同车型客户忠诚度与综合满意度的相关系数

	客户忠诚度（%）	车型 A（%）	车型 B（%）	车型 C（%）	车型 D（%）	综合满意度（%）
1 月	80.4%	20.0%	96.7%	83.7%	88.0%	98
2 月	71.0%	64.7%	90.9%	85.3%	43.1%	100
3 月	69.4%	50.0%	83.3%	81.1%	55.4%	88
4 月	78.4%	21.4%	76.9%	91.3%	87.9%	64
5 月	66.1%	46.7%	84.0%	87.5%	52.9%	82
6 月	81.7%	100.0%	71.4%	75.0%	91.3%	72
7 月	83.3%	36.4%	91.7%	87.5%	85.3%	98
8 月	69.1%	85.7%	90.0%	90.9%	40.4%	68
9 月	70.1%	64.3%	81.6%	72.4%	60.4%	90
10 月	82.2%	40.0%	90.3%	85.7%	78.2%	89
11 月	62.8%	23.1%	80.0%	77.8%	47.1%	91
12 月	64.9%	50.0%	65.7%	60.6%	76.5%	97
1 月	66.9%	50.0%	74.4%	61.3%	74.7%	75
2 月	61.3%	50.0%	62.5%	69.8%	50.4%	65
3 月	68.6%	44.4%	62.5%	64.9%	87.5%	55
4 月	68.4%	48.0%	64.3%	69.2%	91.7%	59
5 月	64.8%	76.9%	48.7%	59.5%	97.9%	69
6 月	53.8%	50.0%	57.1%	51.2%	55.5%	33
7 月	55.2%	26.7%	51.1%	61.1%	84.7%	48
8 月	67.3%	90.9%	66.7%	61.8%	75.4%	48
9 月	55.1%	26.3%	62.2%	50.0%	68.5%	55
10 月	54.8%	33.3%	60.4%	54.8%	60.0%	57
11 月	52.7%	20.0%	66.7%	50.0%	64.4%	44
12 月	59.3%	55.6%	56.7%	52.6%	93.3%	31
与综合满意度的相关系数	0.64	−0.02	0.75	0.69	−0.19	

图 C （车型 B）各项要素与综合满意度的相关

综合满意度	售后服务满意度	产品满意度	同类产品价格比
98	98	75	0.95
100	95	85	0.81
88	50	95	0.96
64	77	100	1.05
82	81	85	1.19
72	62	30	1.04
98	70	55	0.93
68	50	40	0.95
90	86	25	1.05
89	48	35	0.98
91	100	40	0.91
97	22	70	0.90
75	40	70	1.29
65	80	60	1.02
55	50	75	1.24
59	42	65	1.30
69	66	70	0.92
33	32	55	1.33
48	55	60	1.07
48	97	70	1.14
55	73	55	1.04
57	96	95	1.19
44	72	40	1.21
31	35	95	1.24
与综合满意度的相关系数	0.26	−0.10	−0.72

图 D （车型 C）各项要素与综合满意度的相关

综合满意度	售后服务满意度	产品满意度	同类产品价格比
98	89	90	1.29
100	89	15	1.08
88	81	25	0.89
64	67	20	0.93
82	72	30	1.19
72	88	90	1.29
98	70	85	0.92
68	47	20	1.29
90	65	100	1.30
89	81	65	1.15
91	68	20	0.89
97	71	55	1.21
75	52	70	1.12
65	60	85	1.00
55	53	85	1.13
59	73	95	1.02
69	58	50	0.92
33	82	90	1.09
48	61	40	1.05
48	44	60	1.20
55	38	80	1.06
57	43	30	1.08
44	47	95	1.06
31	39	15	1.05
与综合满意度的相关系数	0.59	-0.07	0.09

通过图 C 可以发现，对车型 B 来说，与同类产品的价格比（相对而言是贵还是便宜）对综合满意度的影响较大。二者的相关

系数为-0.72，表示价格越高，顾客满意度就会越低。需要注意其变化趋势是相反的，也就是说，车型 B 的用户对价格比较敏感。

对车型 C 也进行同样的调查，结果为图 D。可知对车型 C 来说，售后服务满意度与综合满意度高度相关（0.59）。

同样是综合满意度，车型 B 的用户与车型 C 的用户所重视的关键点完全不同。当然，我们也可以越过综合满意度，直接考察每个车型的客户忠诚度与“售后服务”“产品”“价格”等数据的相关系数。

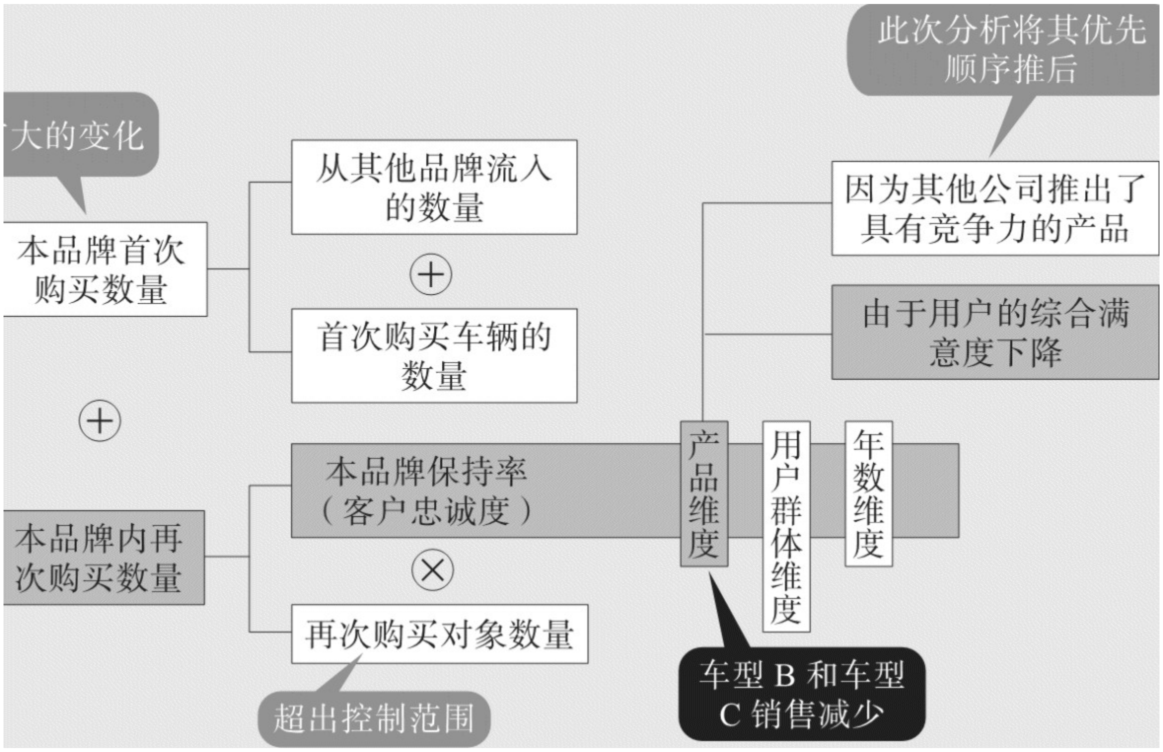
不过在提出假设时，直接连接到具体项目，就有产生“逻辑跳越”的风险，可能会有人质疑：“客户忠诚度为什么会与售后服务直接联系在一起”。为了保证自己的故事能够令人信服地说明整个经过，需要细心地构建假设，依据假设依次分析，这会左右整个分析的可信度。

将前面所有分析组织起来，其结构如图 E 所示。它体现了对问题进行深入挖掘的整个过程。

要注意，必须确保沿着挖掘出的原因逆流而上，一定会达到“销售总额减少”这个最根本的问题。

此外，图 E 还体现出，分析者并不是只分析了偶然想到的某些项目，而是通过这个构造避免了遗漏或重复，并对那些最终确定并非问题或原因的项目也进行了检验。另外，对话框里的内容解释了停止深入挖掘的原因。这样一来，听众就可以明白，分析者说到解决问题、进行分析或企划的目标时，“该工作的最终目的”这一重要视点始终没有动摇。锁定原因之后，接下来就是制定改进（解决）措施了。

图 E 分析得出的解决问题的故事的构造



大的变化
本品牌首次购买数量
从其他品牌流入的数量
首次购买车辆的数量
本品牌内再次购买数量
本品牌保持率（客户忠诚度）
再次购买对象数量
超出控制范围
产品维度
用户群体维度
年数维度
车型 B 和车型 C 销售减少
因为其他公司推出了具有竞争力的产品
此次分析将其优先顺序推后
由于用户的综合满意度下降

第 4 章
制定对策，要依据“方程式”

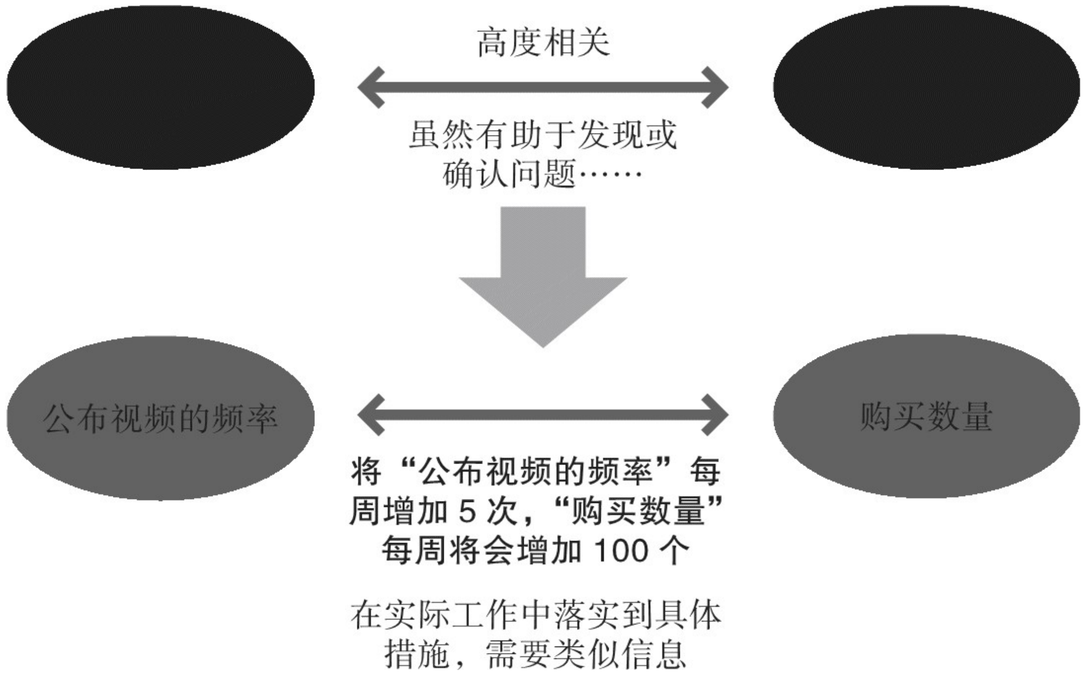
前面介绍了如何确认 2 个数据之间的相关程度并锁定原因。相关分析具有简单实用的优点，即使数据单位不同［例如“人”和“钱（日元）”等］，也可以进行分析。但在实际工作中，要想充分发挥数据分析的作用，获得对方的认可，还需更进一步的工作。

例如，相关分析的结果表明，“在网上公布视频的频率”与“购买数量”高度相关（图 4-1）。根据数据得出这一发现虽然有意义，但设想一下，如果把“高度相关”作为结论汇报给上司，他会做出怎样的反应呢?

上司一定会问：“我知道它们高度相关了。那么，公布视频的频率对购买数量到底有多大影响呢？”了解相关程度的大小，对锁定原因非常有效，但只靠相关分析却无法得知这个原因会对目标产生多大影响。

虽然找到了高度相关的原因，却仍然不知道要改善到什么程度才能对目标带来变化。而对掌管业务整体运营的人来说，这是一个非常重要的问题。

图 4-1 公布视频的频率与购买数量的关系



此外，上司可能还会问及措施或计划：“那么具体要采取哪些措施，做到何种程度呢？”也就是说，相关分析归根结底还只是锁定原因，而无法规划下一步的措施。

现实工作中，人们需要根据数值采取行动。“为什么要进行分析”“计划根据分析结果做什么”，这些视点和思维方式，对实际业务中的数据分析工作具有极为重要的意义。

因此，继相关分析之后，还需要“一元回归分析”登场。一元回归分析可以将 2 个数据之间的相关关系表现为具体公式。前文的例子就可以通过一元回归分析计算出“在网络公布视频的频率”为多少时，“购买数量”会达到多少。

10 秒钟完成一元回归分析

例如，如图 4-2 所示，假设有 30 个星期内每周公布视频的频率和购买数量的数据。首先使用 CORREL 函数确认二者之间的关联，得出相关系数为 0.80，属于高度相关。接下来，可以按照以下步骤，得出“公布视频的频率”与“购买数量”的数值关系。

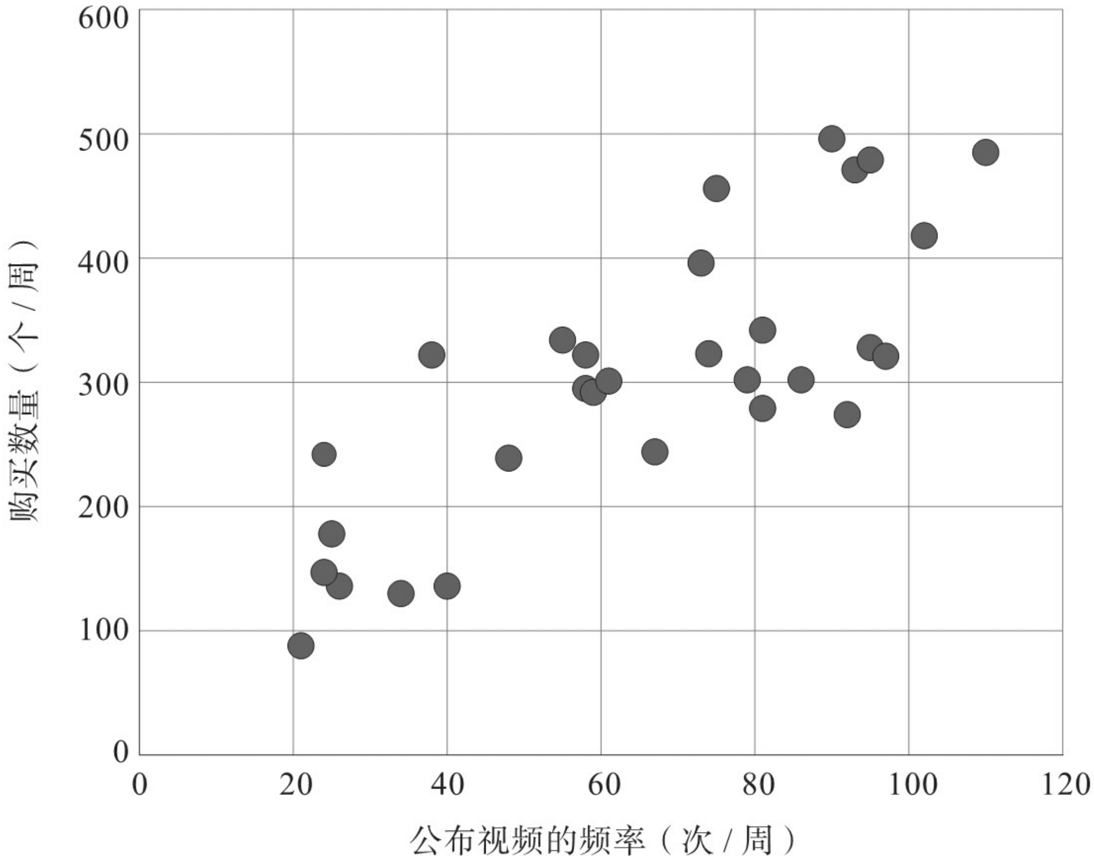
图 4-2 30 个星期内每周公布视频的频率和购买数量的数据

周	公布视频的频率（次 / 周）	购买数量（个 / 周）
1	95	328
2	75	456
3	110	485
4	86	302
5	73	396
6	24	242
…	…	…
27	38	322
28	97	321
29	81	342
30	34	130
	相关系数	0.80

（1）用散点图展现 2 个数据间的关系

绘制散点图时，一般需要注意确认将哪个数据设为纵轴（图 4-3）。除了清晰明了、简单易懂之外，散点图还必须保证在之后的数值化过程中，用纵轴表示“输出（结果或目的）”，用横轴表示“输入（能够控制的变量）”。散点图的原则是通过控制横轴的变量，引起纵轴变化。在 Excel 中将纵轴的数据放在右侧，横轴的数据放在左侧，就可以得到想要的散点图。

图 4-3 公布视频的频率与购买数量的散点图



（2）用散点图求回归方程

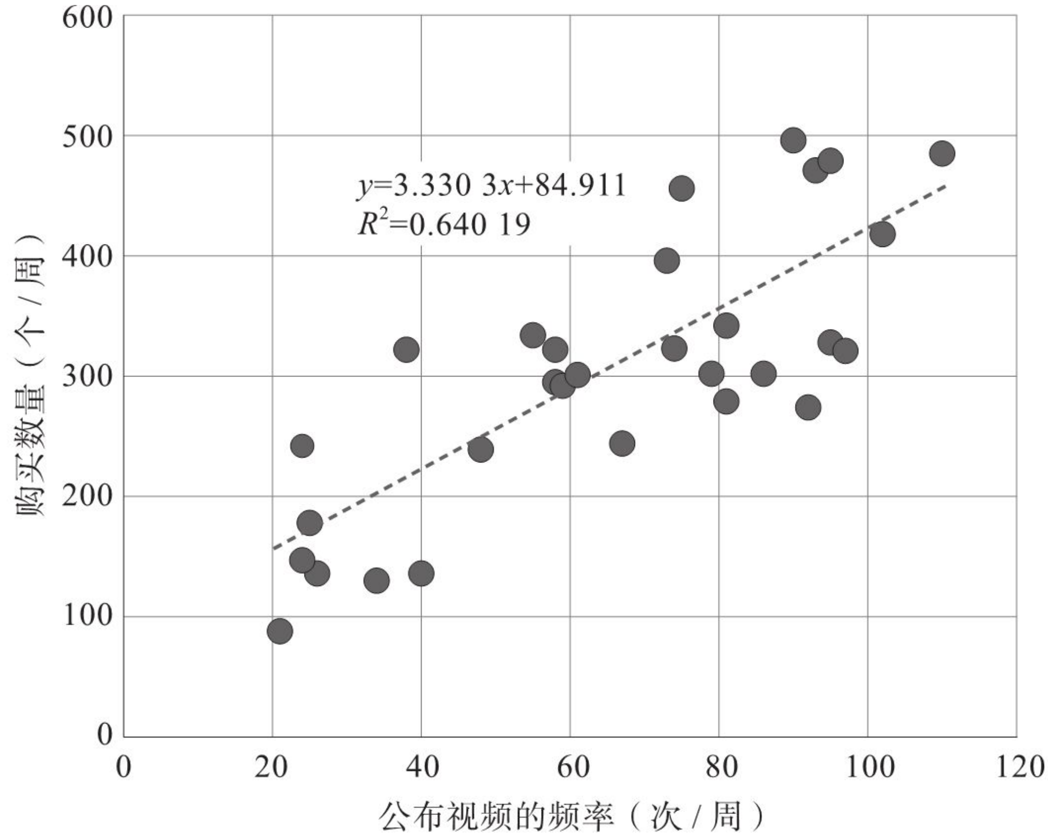
选择散点图上任意一个点，点击鼠标右键，在菜单中选择“添加趋势线”。然后在图 4-4 所示的页面中勾选最下面的“显示公式”和“显示*R*平方值”。如图 4-5 所示，散点图上就会出现一条大致从数据中心通过的直线，以及体现纵轴与横轴数值关系的公式。

图 4-4 勾选在图表中显示公式和*R*平方值



依据具有相关关系的数据，上述操作可以在 10 秒钟之内完成。这个方法与相关分析一样，能在实际工作中发挥巨大优势。在实际工作中，我们收集到的数据，并不一定都能预测出结果，或者说可能大多数都无法预测结果，所以需要在较短时间里多次试错。能在 10 秒钟之内完成一次分析，也就意味着可以在有限的时间里进行多次检验。

图 4-5 散点图上的回归方程和R 平方值（R^2）



那么，应该如何解释这个结果呢？散点图上的公式为：

y=3.330 3x+84.911

在这个例子中，该公式可以理解为：

购买数量（个/周）=3.33×公布视频的频率（次/周）+84.9

可能很多人都发现了，我们曾经在中学数学课上学过这个公式，表示直线$y=ax+b$（a是斜率，b是常数）。这条直线就是图 4-5 中的直线。一元回归分析公式用$y=ax+b$的形式来表现这条直线，这就是回归方程（由此画出的直线叫作回归直线）。

回归直线作为零散分布的原始数据的代表，是距离各点（数据）之和最小的直线。不过除非所有数据都排列在同一条直线上，否则回归方

程与各点之间就一定会有偏差。就这一点而言，回归方程无法完美地体现原始数据。

这样一来，就还需要另一个指标，来衡量回归方程（作为原始数据的代表）的可信度。这个指标就是散点图里写在回归方程下面的R^2。

让我们再来看看相关分析。如果原始数据完全相关（相关系数=1），那么所有的点都会排列在一条直线上；但如果不是完全相关，相关系数就会随着数据对完全相关（直线）的偏离，从 1 开始逐渐减小。其实，越偏离直线，指标越小于 1 的现象也适用于一元回归。数据偏离越远，回归直线就越不能准确地代表原始数据，R^2表示数据的偏离程度。

关注相关系数的平方

接下来，我们来看R^2。在图 4-2 的例子当中，相关系数是 0.80。其平方为 0.64，与从散点图求出的R^2 =0.640 2 的数值一致。R^2就是相关系数的平方。因此我们对R^2也可以采取与相关系数相同的评价标准。

多数情况下，我们将相关系数大于 0.7，或者稍微放宽一些，将相关系数大于 0.5 的情况视为“相关”。0.7 和 0.5 的平方分别为 0.49 和 0.25。大多数情况下，我将 0.7 作为相关系数的标准，将 0.49 作为R^2的标准，大于这个数值则判断数据相关，可以放心地使用回归方程。对同一数据来说，根据相关系数判断，还是根据R^2判断，其结果是相同的。

顺便说一下，x叫作“自变量”或“解释变量”，y叫作“因变量”或“被解释变量”。不知道这些专用名词也不会影响分析，不过与了解回归分析的人交流时，自然会用到这些词，做些了解也没有坏处。

如果用百分数（%）表示R^2，它可以理解为“被解释变量”在多大程度上可以由“解释变量”来说明。就前文这个事例来说，就是在“购买数

量”这个变量当中，有 64.02%（=0.640 2）可以通过“公布视频的频率”得到解释。这样可能更便于我们理解R^2的含义。

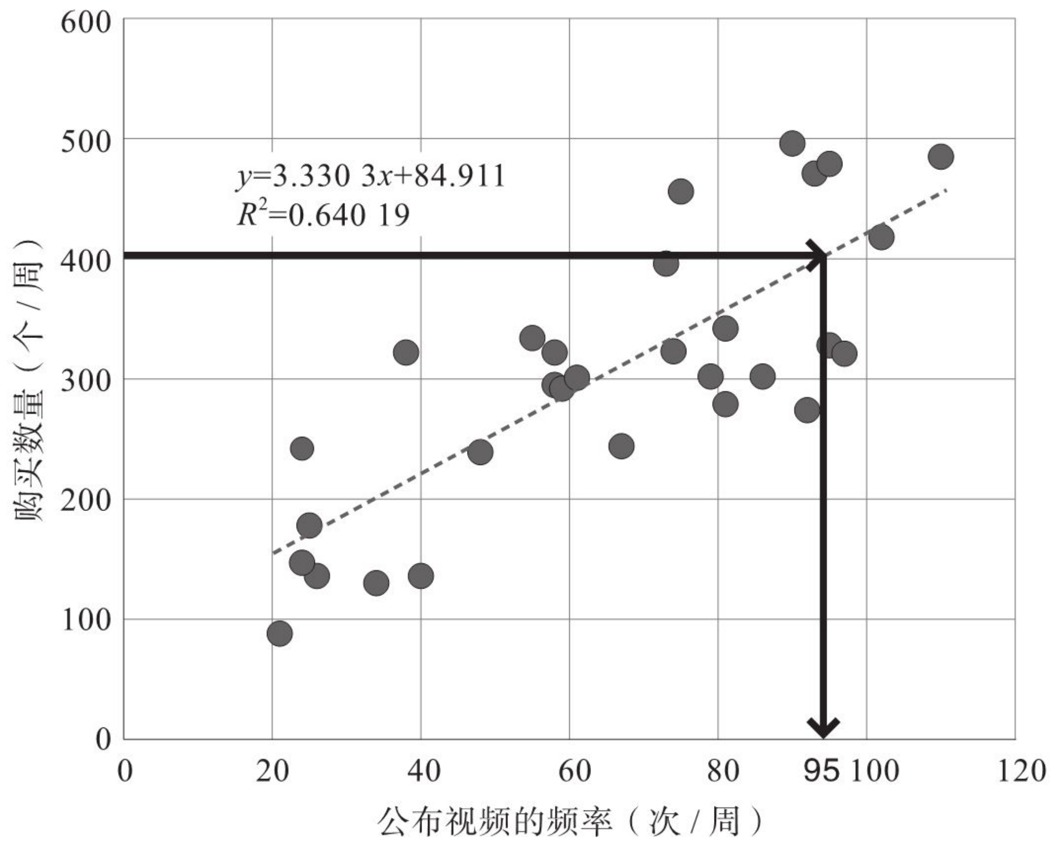
那么，刚才得出的回归方程，应该如何解释呢？斜率（3.33）表示，“公布视频的频率”每周增加 1 次，“购买数量”将会增加 3.33 个。比较公布视频的频率增加 1 次的成本和购买数量增加 3.33 个所带来的收益，就可以检验目前公布视频的成本能否带来足够的销售增长。如果公布视频 1 次需要 1 000 日元的成本，而销售增加 3.33 个带来的利润增长为 900 日元的话，我们就不会再进一步增加公布视频的频率。

只靠相关分析的结果无法得到类似的数值关系。通过这种关系可以得知，增加 1 次公布视频的频率，会对目标即购买数量产生多大的影响。如果其他要素与购买数量的相关程度较低，但能给购买数量带来更大影响的话，也可以优先采取那个方面的措施。

另一方面，假设本周销售目标为 400 个，也可以将它代入y，反向推算公布视频的频率（400=3.33×公布视频的频率+84.9）。计算可知，需要公布视频的频率为 95 次/周。将这种情况反映在散点图上就是图 4-6。

像这样，知道达到目标（例如销售数量 400 个）所需要的输入（例如公布视频的频率）为多少，就可以计算出需要多少资源，或者据此设定行动指标（KPI：重要业绩评价指标），从而制定出更为客观并符合逻辑的计划。

图 4-6 用散点图逆向推算公布视频的频率



后文还会专门介绍具体应用事例，希望这种方法能帮助大家告别只靠"毅力和热情"来制定计划的做法。当然，也可以先设定公布视频的频率，再根据回归方程计算能带来多少购买数量。

注意事项及应用事例

运用一元回归分析要注意以下两个问题。

一个是"数据之间必须具有单纯的比例关系"。与相关分析一样，一元回归分析的大前提是 2 个数据之间存在直线比例关系。情况或关系越复杂，就越不符合严密的线性关系。这种情况就不适合套用$y=ax+b$。

Excel也有计算曲线回归方程的功能。从理论上讲，我们可以对两种方式取得的R^2值进行比较，选择R^2值较高的方式。但考虑到在实际业务中的运用，即使R^2值略低，但只要高于一定标准（例如0.25或0.49），还是建议使用一元回归分析。

因为一元回归分析除了计算简单之外，在说明分析过程时，也更容易得到非专业人士的理解。如果对方无法理解分析过程或所使用的理论，一般也就无法接受由此得出的结论，最终就无法达成共识。实际工作中的难点其实就在这里。

还有一个需注意的问题，即“离群值以及不同的数据选择范围，会导致分析结果产生很大不同”。这一点也与相关分析的注意事项一样。是否采纳明显偏离其他数据的“离群值”，如何设定数据的范围（例如是过去半年期间的数据，还是一年期间的数据），分析者的不同判断会导致分析结果出现很大差异。也就是说，分析者每一个小小的决定都能操纵分析的结果。

当然，分析者必须要对离群值及数据范围等的处理方法加以说明。把手边所有的数据都用上，边做边看能得到什么结果的做法，无论准确度还是可靠性都无法令人信服。

一元回归分析是一种极其卓越的方法，只用10秒钟的时间就可以使用Excel计算出数据间的关系（回归方程或回归直线）。但工作中进行数据分析的最终目标并不是用数字来表示数据关系。只有数据关系得到了充分的运用，分析才具有价值。

那么，该如何运用呢？几乎所有的数据分析教科书都没有涉及。这也是很多人在实际工作中最大的烦恼。不同的业务、行业或者商业形态，具有无数种运用数据关系的模式。希望下面介绍的3个事例能为大家带来启发，成为大家在工作中应用数据分析时的参考。

事例1　对比成本和收益

“把有限的资源（时间、成本）投入到哪里、投入多少才能发挥最大效率”“现行的成本使用方式是否真能带来预期的收益”……尽管我们常有这些疑问，一般却不会去检验，而是任由资源浪费。那么接下来，我们就尽量具体地考虑一下，考察哪些关系，才能用数字体现成本与收益的关系。

计算“成本”和“收益”的数值关系，需要根据问题内容，确定具体是哪些指标（数据）代表成本与收益，否则就不知道应该使用哪些数据。首先看成本，我们需要定义它具体指哪些内容。例如是只有广告宣传费，还是也包括相关的人工费在内等。

对于收益，同样也需要明确此处所说的收益是用哪些数据衡量的。例如，除了常用的“销售额”“销售额增长率”之外，还可以考虑“来店人数”“咨询件数”等指标。

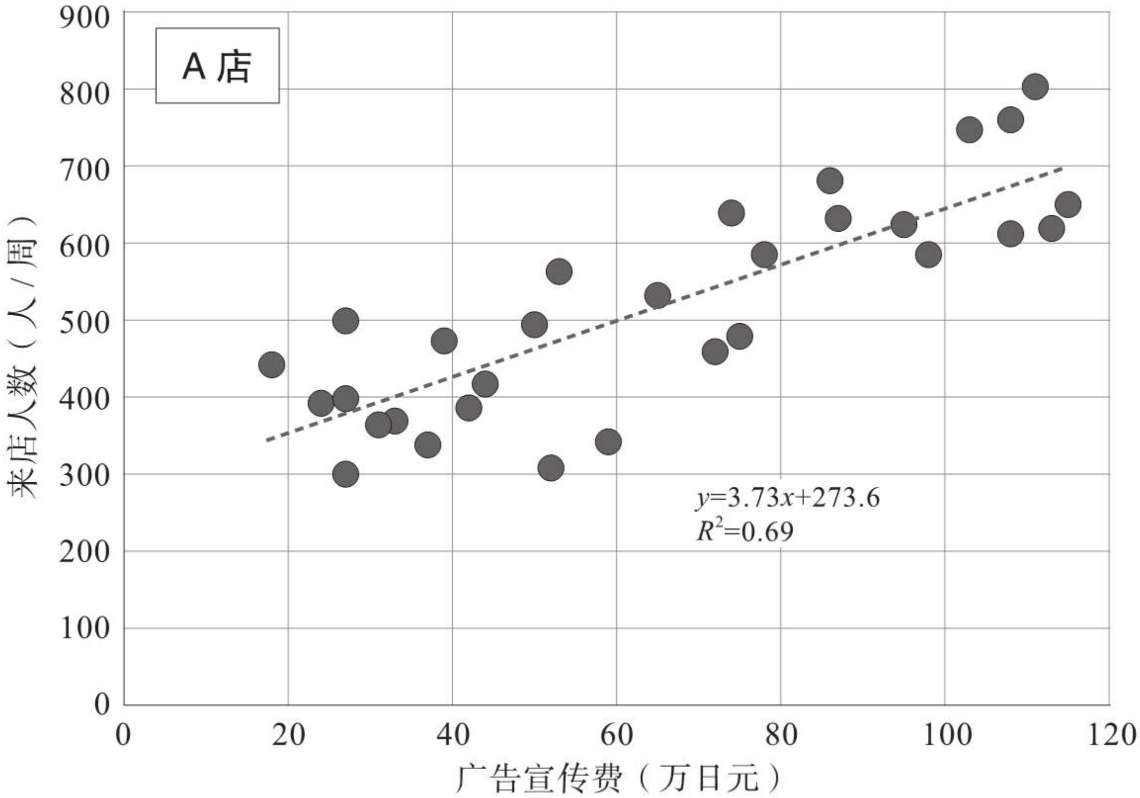
哪项指标最适合用来分析，取决于它能否充分体现出分析的目的，而且除了要考虑这种数据是否能收集到，还有一个前提是它与成本之间必须存在相关关系。图 4-7 以“广告宣传费”作为成本，以“来店人数”作为收益，对 A 店和 B 店的成本和收益进行了比较。

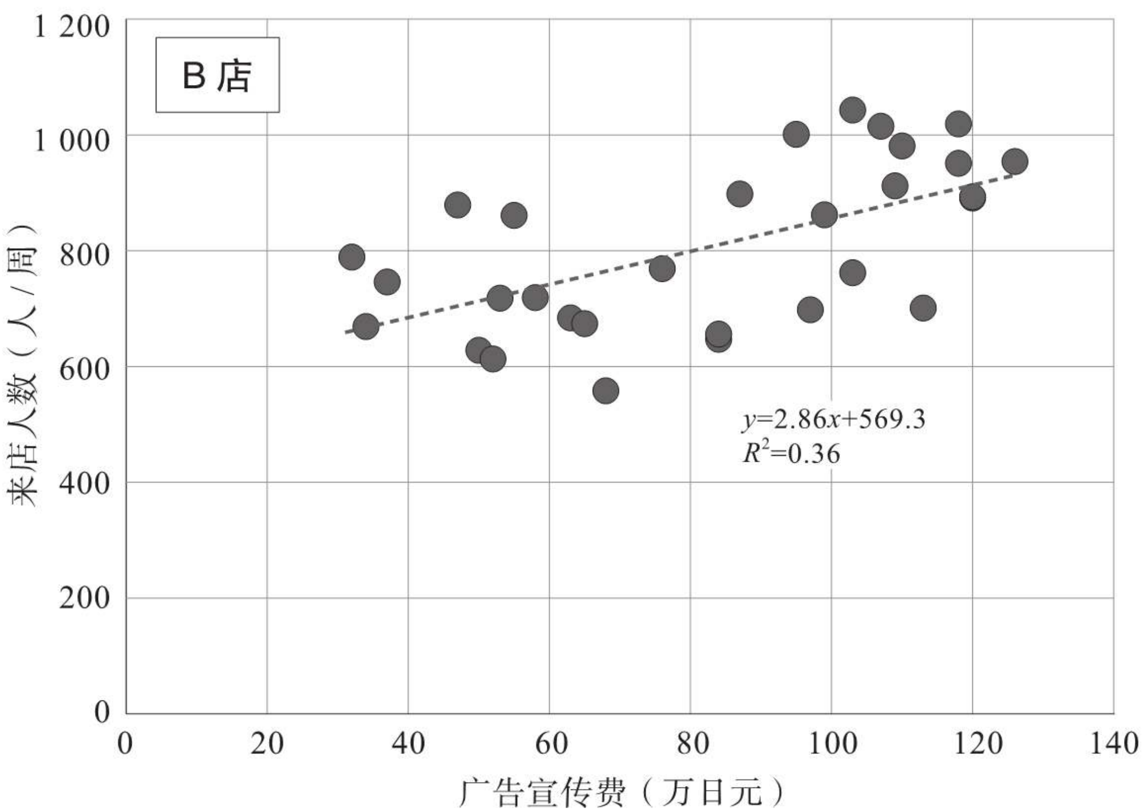
如果不用数值表示，而是直接站在店里观察来店的人数，会怎样呢？假设在使用了 60 万日元广告宣传费的那一周，A 店观测到的来店人数约为 500 人，B 店约为 750 人。B 店的绝对数值更大，似乎可以得出“B 店收益更高”的结论。只看来店人数多少的话确实如此，但是因为广告宣传而来店的人数和并非因为广告宣传而来店的人数是混在一起的，仅靠这一点并不能得知广告宣传（费）带来的真正收益。

A 店的回归方程如下：

来店人数（1 周的人数）=3.73×广告宣传费（万日元）+273.6

图 4-7 广告宣传费与来店人数的关系


A 店
来店人数（人 / 周）
广告宣传费（万日元）
y=3.73x+273.6
R2=0.69
900
800
700
600
500
400
300
200
100
0
0
20
40
60
80
100
120


B 店
来店人数（人 / 周）
广告宣传费（万日元）
y=2.86x+569.3
R2=0.36
1 200
1 000
800
600
400
200
0
0
20
40
60
80
100
120
140

请大家注意，该方程的斜率为 3.73，这意味着广告宣传费每增加 1 万日元，来店人数可以增加 3.73 人。这个数值代表 1 万日元能够带来多大的收益。同样可知，B 店每增加 1 万日元的广告宣传费，来店人数会增加 2.86 人。同样是 1 万日元，用在哪一家店铺的效果更好（即成本带来的收益更高）呢？比较斜率，3.73 大于 2.86，因此 A 店的收益更高。

这个思维方式在前文“购买数量与公布视频的频率”的事例中曾经介绍过，它在锁定问题原因时也可以发挥作用。针对某个问题，如果存在多个具有相关关系的原因，就需要讨论这些原因之间的优先顺序。其中一个做法是从“关联更为密切”的角度，比较哪个原因与问题的相关程度更高。

另一方面，根据回归分析的结果，可以得知某个原因能够对问题造成“多大程度”的影响。回归方程的斜率体现了改善解释变量的成本，与问题由此得到改进的关系。因此可以根据斜率，从“成本和收益”的角度决定优先顺序。

看相关系数还是看斜率

经常有人问我，“应该按照相关系数判断，还是按照成本和收益的比较结果判断？”我一般会像下面这样回答。

“相关系数常作为筛选标准，与 0.7、0.5 等界限值进行比较，只要高于界限值就可以判断是相关。由于其中会包含误差等，所以并不按照其具体大小进行判断。对已经确定具有相关关系的数据，则可以运用回归分析来定量地比较或评价其影响。也就是说，可以从 2 个视角来看，用相关分析判断关联的紧密程度，用回归分析判断其影响大小。”

下面，为了进一步理解回归分析的结果，我们再来思考，同样使用了 60 万日元的广告宣传费，为什么收益低的 B 店的来店人数会更多。

其原因在于，来店人数中包含了并非因为广告宣传而来店的顾客。我们可以从“理论上”计算出不使用广告宣传费（即 0 日元）也会来店的人数。在回归方程中，把 0 代入广告宣传费，其结果就是无论是否进行广告宣传都会来店的人数。

A 店：来店人数（人）=3.73×0+273.6=273.6（人）

B 店：来店人数（人）=2.86×0+569.3=569.3（人）

也就是说，即使没有广告，B 店仍然会有约 570 人来店，远远多于 A 店的约 270 人。这可能是店铺选址或商圈等方面的差异造成的。从使用 60 万日元广告宣传费时的来店人数（A 店约 500 人、B 店约 750 人）中，减去并非因为广告宣传而来店的人数：

A 店：500 -约 270 =约 230（人）

B 店：750 -约 570 =约 180（人）

可知，使用 60 万日元广告宣传费，A 店能够吸引来的人数比 B 店多约 50 人。在现实工作中，上述内容不进行零广告宣传费的试验，就无法得知。而运用回归分析，就可以轻松地计算出理论上的数值结果。这正是回归分析的魅力之一。

那么将 B 店的广告宣传支出削减为零，把所有费用都集中到 A 店会更好吗？当然没有这么简单。无限增加广告宣传费，来店人数就会成比例地无限增长，这种想法是不现实的。回归分析归根结底只是在所使用的数据范围内有效。而且不要忘了，正如前文介绍的，一元回归分析只是将现实世界嵌套在极其单纯“比例关系”模式中。它非常易懂易用，但也因为过于简化而存在缺陷。分析者应该了解这一点。

事例 2　合理分配资源

让我们再次来看 A 店的例子。假设 A 店根据来店人数的目标，计算出所需广告宣传费，并希望将该金额反映在年度预算计划中。例如，配合某月新商品上市，为了吸引更多的顾客来店，将目标设定为 3 200 人。单纯按照每月 4 个星期计算，则相当于每周 800 人。为了实现 800 人的目标，需要的广告宣传费如下：

800（人）= 3.73 ×广告宣传费（万日元）+ 273.6

根据这个方程式计算，得知每周所需的广告宣传费约为 141 万日元，所以这个月（4 周）需要的预算约为 560 万日元。与缺乏可靠根据的预估金额相比，这种方法能够明确地体现出为达到什么样的数值目标，需要多少预算。这样不仅能够提高预算金额的准确度，提交高层管理者或上司审批时也更具说服力。

事例 3　设定合理的 KPI

很多公司或者组织只是为了完成眼前的任务而努力，对“工作的最终目标是什么”“如何评价工作的结果”等根本性问题都没有明确的定义。如果把“看上去是否努力”作为判断产出的晴雨表，除了会导致加班费无限增加之外，还会让员工看不到目标而盲目前进。这样的状态下，公司是无法取得可靠业绩的。

因此，KPI 受到了人们的广泛关注。不过很多组织虽然已经意识到 KPI 的重要性，并将其导入工作中，但实际上，很多 KPI 是随意决定的，或者只是对上一年度业绩稍做调整。

在下面的事例中，某设施为了增加使用人数（提高使用率），针对相关分析得出的原因，设定了理论上的 KPI。假设使用人数与使用满意度之间存在相关关系。

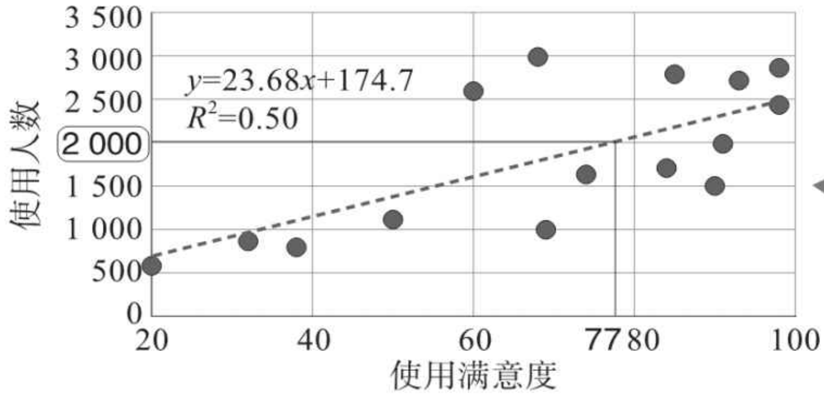
如图 4-8 上方的散点图所示，假设下一年度使用人数的目标为 2 000 人（该目标可以根据维持设施所需的最少人数等计算，也可以根据经营计划、中期计划等目标来设定）。

根据过去数据进行回归分析，可以倒推出使用人数达到 2 000 人所需达到的"使用满意度"。该事例可以得出以下回归方程式：

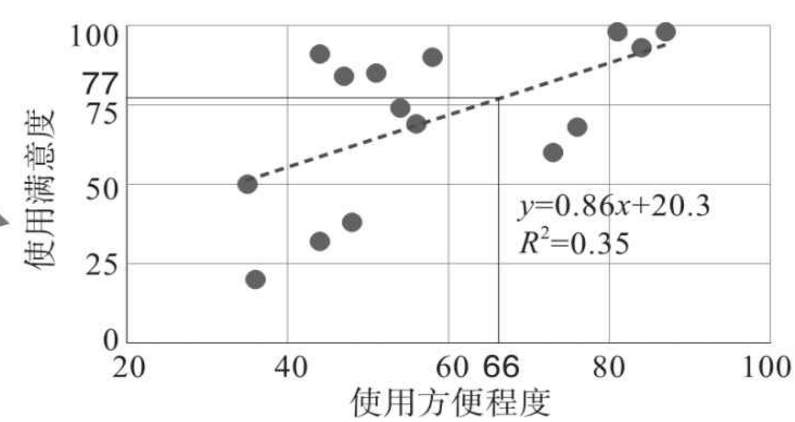
2 000（人）= 23.68×使用满意度（分）+ 174.7

计算可知，需要达到的使用满意度约为 77 分。从散点图上也可以确认到这个结果。那么，怎样才能让使用满意度达到 77 分呢？只提出"提高使用满意度"的口号，员工还是无法采取具体行动。于是接下来关注与试用满意度高度相关的"使用方便程度"，它是提高使用满意度的更进一步原因。然后像图 4-8 下面的散点图一样，对"使用满意度"与"使用方便程度"进行回归分析。与之前一样，可以用回归方程式反向计算出，使用满意度要达到 77 分，使用方便程度的分数需要达到约 66 分。可见，要实现 2 000 人的使用人数目标，使用满意度需要达到 77 分，为此需要将使用方便程度提高到至少 66 分。采用这个方法，便可以将

图 4-8 通过 KPI 实现使用人数目标



	使用方便程度（100 分满分）	使用满意度（100 分）	使用人数（人 / 月）
设施 A	87	98	2 434
设施 B	48	38	794
设施 C	73	60	2 593
设施 D	81	98	2 861
设施 E	84	96	2 715
设施 F	44	32	863
设施 G	76	68	2 987
设施 H	36	20	581
设施 I	35	50	114
设施 J	44	94	1 986
设施 K	58	90	1 501
设施 L	56	69	997
设施 M	54	74	1 633
设施 N	47	84	1 707
设施 O	51	85	2 788



KPI 分别设定为 77 分和 66 分。接下来便可以每月进行问卷调查，参考目标值，检测目前所处的“位置”，从而采取适当的改进措施。

当然，提高使用满意度的方法不仅限于提升使用方便程度（正因为如此，R^2值也不是 1）。而且事实上，设定 KPI 的这个方法说到底不过是根据过去数据得出的理论数值。

不过，希望读者能把这个方法与完全没有任何指导方针、不设定目标终点、只是拼命努力的情况，或者毫无根据地把目标数值强加给一线工作人员的情况做一个比较。两者在进度管理的有效性、一线员工的接受程度等方面，应该都有很大差别。

我在日产工作时，曾经很多次绘制回归直线，比较斜率，从而考察某个行动（散点图上的横轴）对某个目标（散点图上的纵轴）来说是否有效，或者检验回归直线是否至少是向右上方倾斜的（即越采取行动越能产生效果）。对各种背景的人来说，这种方法都能够一目了然地展现出某个行动是否有效。

不过在现实中，也有很多情况无法顺利求出R^2值或相关系数，此时就比较容易出现意见分歧。虽然不一定 100% 有效，不过可以将分析对象划分为不同的地区或国家，或者划分成多个期间分别尝试，有时便能在某个范围内找到更为详细（不同情况下）的答案。由此出发，着眼于该范围（不同地区等）特有的问题，就能够对问题进行深入挖掘。

「解决问题的故事 4」

第四步：通过一元回归分析，发现车型 B 和车型 C 的不同问题

前面通过相关分析，已经得知车型 B 和车型 C 的问题原因分别为“相对价格”和“售后服务”。接下来考虑如何解决车型 B 的问

题。

车型 B 的客户忠诚度在过去 6 个月里跌至约 60%。虽然 1 年多以前的客户忠诚度接近 90%，但马上恢复到当时的水平是不现实的，因此公司决策层提出的要求是，在 6 个月以内提高 25%，即将客户忠诚度恢复到 75%（60×1.25）。

如何使用数字将客户忠诚度 75% 的目标与一线作为行动目标的 KPI 联系在一起，这关系到后面的工作能否取得效果。

由于车型 B 的客户忠诚度与综合满意度之间的相关系数为 0.75，属于高度相关，因此可以运用一元回归分析考察二者之间的数值关系（图 A）。根据一元回归分析得出的回归方程式，可以计算出达 75% 的客户忠诚度所需的综合满意度分数。

图 A 客户忠诚度与车型 B 的满意度数据

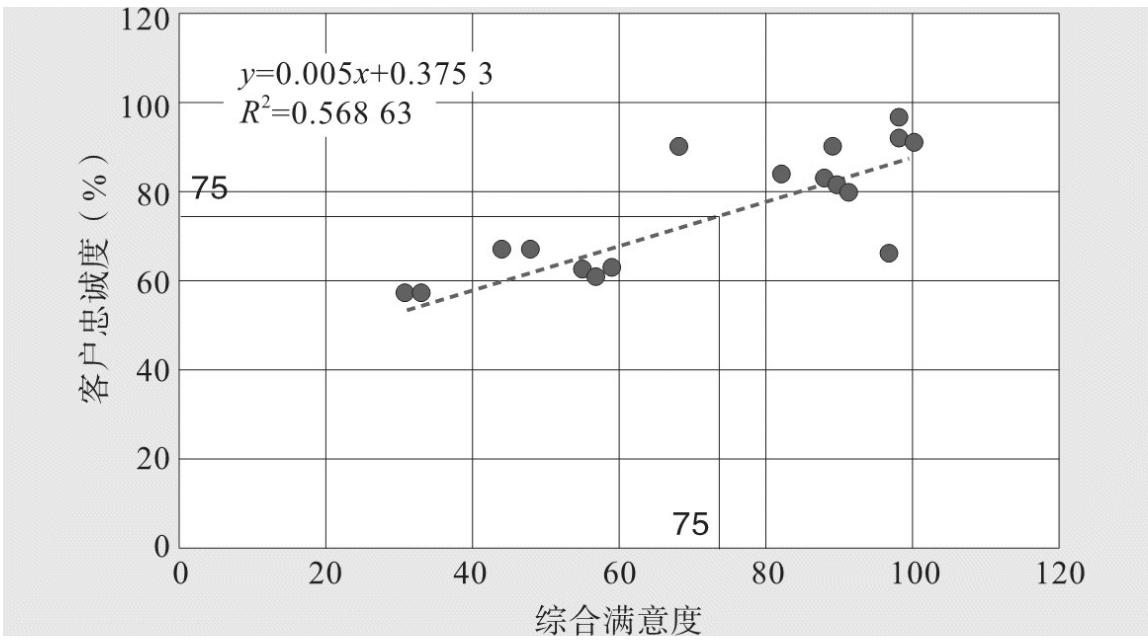
	客户忠诚度（%）	车型 A（%）	车型 B（%）	车型 C（%）	车型 D（%）	综合满意度	售后服务满意度	产品满意度	同类产品价格比
1月	80.4	20.0	96.7	83.7	88.0	98	98	75	0.95
2月	71.0	64.7	90.9	85.3	43.1	100	95	85	0.81
3月	69.4	50.0	83.3	81.1	55.4	88	50	95	0.96
4月	78.4	21.4	76.9	91.3	87.9	64	77	100	1.05
5月	66.1	46.7	84.0	87.5	52.9	82	81	85	1.19
6月	81.7	100.0	71.4	75.0	91.3	72	62	30	1.04
7月	83.3	36.4	91.7	87.5	85.3	98	70	55	0.93
8月	69.1	85.7	90.0	90.9	40.4	68	50	40	0.95
9月	70.1	64.3	81.6	72.4	60.4	90	86	25	1.05
10月	82.2	40.0	90.3	85.7	78.2	89	48	35	0.98
11月	62.8	23.1	80.0	77.8	47.1	91	100	40	0.91
12月	64.9	50.0	65.7	60.6	76.5	97	22	70	0.90
1月	66.9	50.0	74.4	61.3	74.7	75	40	70	1.29
2月	61.3	50.0	62.5	69.8	50.4	65	80	60	1.02
3月	68.6	44.4	62.5	64.9	87.5	55	50	75	1.24
4月	68.4	48.0	64.3	69.2	91.7	59	42	65	1.30
5月	64.8	76.9	48.7	59.5	97.9	69	66	70	0.92
6月	53.8	50.0	57.1	51.2	55.5	33	32	55	1.33
7月	55.2	26.7	51.1	61.1	84.7	48	55	60	1.07
8月	67.3	90.9	66.7	61.8	75.4	48	97	70	1.14
9月	55.1	26.3	62.2	50.0	68.5	55	73	55	1.04
10月	54.8	33.3	60.4	54.8	60.0	57	96	95	1.19
11月	52.7	20.0	66.7	50.0	64.4	44	72	40	1.21
12月	59.3	55.6	56.7	52.6	93.3	31	35	95	1.24
与综合满意度的相关系数	0.64	−0.02	0.75	0.69	−0.19		0.26	−0.10	−0.72

客户忠诚度（75%）= 0.005×综合满意度+ 0.375 3

用这个方程计算综合满意度，约为 75（分）。

根据图 B 的回归直线，也可确认要实现“客户忠诚度 75%”，综合满意度需要达到 75 分。但是只有“使综合满意度达到 75 分”的目标，仍然无法得知“具体要将哪一方面改进到何种程度”。为此，还需要进一步落实到具体内容上。

图 B 综合满意度与客户忠诚度的关系（车型 B）



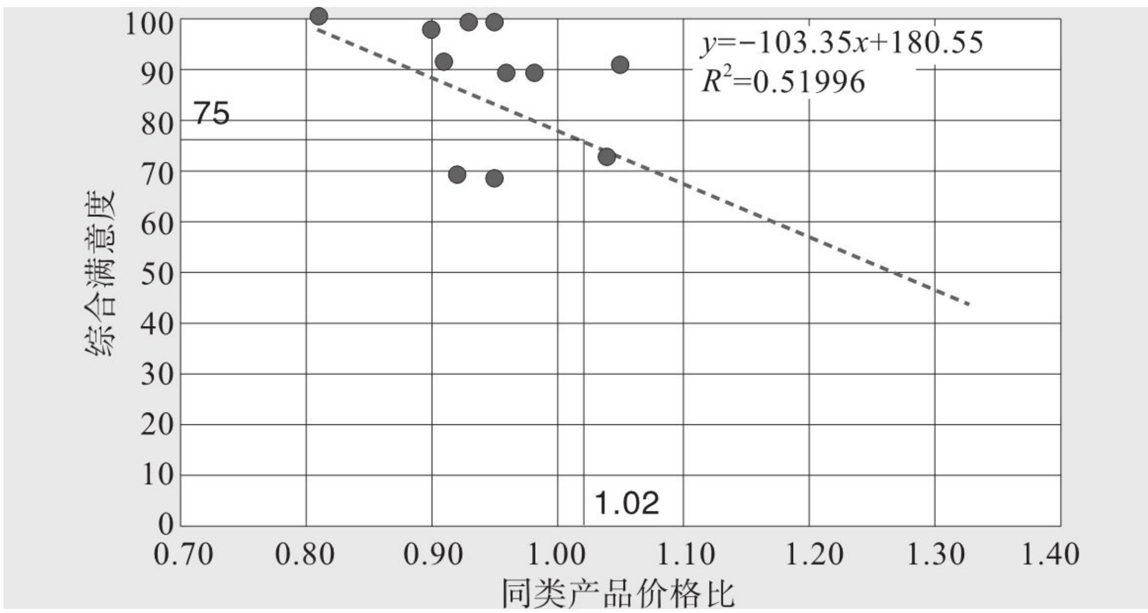
通过分析原因，我们已经得知车型 B 的综合满意度与“同类产品价格比”高度相关。也就是说，车型 B 的用户对该产品与其他公司产品的价格差较为敏感，会影响满意度。

因此，对“综合满意度”与“同类产品价格比”进行一元回归分析，结果如图 C 所示。根据这个回归方程，可以计算出达到 75 分的综合满意度，需要将同类产品价格比维持在何种程度（与之前的思路相同）。

综合满意度（75 分）= -103.35×同类产品价格比+180.55

计算得出的同类产品价格比为 1.02，即车型 B 与同类产品相比，价格高出 2% 之内属于容许范围，如果超过这个范围，从理论上看，综合满意度就很难达到 75 分。

图 C 综合满意度与同类产品价格比的关系



这样就可以采取措施，要求卖场的营业人员、销售公司随时监控同类产品的价格动向，为产品 B 设定同类产品价格+2% 的价格界限值。然后，销售一线就可以针对价格、综合满意度等指标设定目标，为实现该目标而努力。

通过前文的分析，已经明确此项措施在逻辑上会直接关系到销售数量和销售总额等上级目标的提高。这样一来就可以通过故事（逻辑）将一线的措施与上级目标联系起来。

当然，还可以对上级目标进行一元回归分析，计算出客户忠诚度提高到何种程度，销售数量会如何变化，最终能达到多少的销售总额。

在实际工作中，很少有单纯凭借一个原因就能完全解决上级问题的情况。无论多么缜密的公式，应用到工作中时都难以保证完全

跟预想的一样。一般都需要一线的监控和调整，才能使操作实现最优化。不过最开始有无理论上的目标和数值指标，会对其效果或效率带来差异。而且在需要提出对策或计划，获得上级批准时，这一点也会影响其说服力的大小。

车型 C 与车型 B 原因不同

接下来再看车型 C。车型 C 也与车型 B 一样，客户忠诚度与综合满意度高度相关，不过其综合满意度与“售后服务满意度”的相关程度较高（图 D）。

首先，根据对客户忠诚度与售后服务满意度进行一元回归分析的结果可知，要达到 75% 的客户忠诚度所需的售后服务满意度如图 E 所示。本来也应该像车型 B 一样，用“综合满意度”为媒介，先分析“客户忠诚度”与“综合满意度”，然后再分析“综合满意度”与“售后服务满意度”，一步步推进，这样更为细致和合乎逻辑。

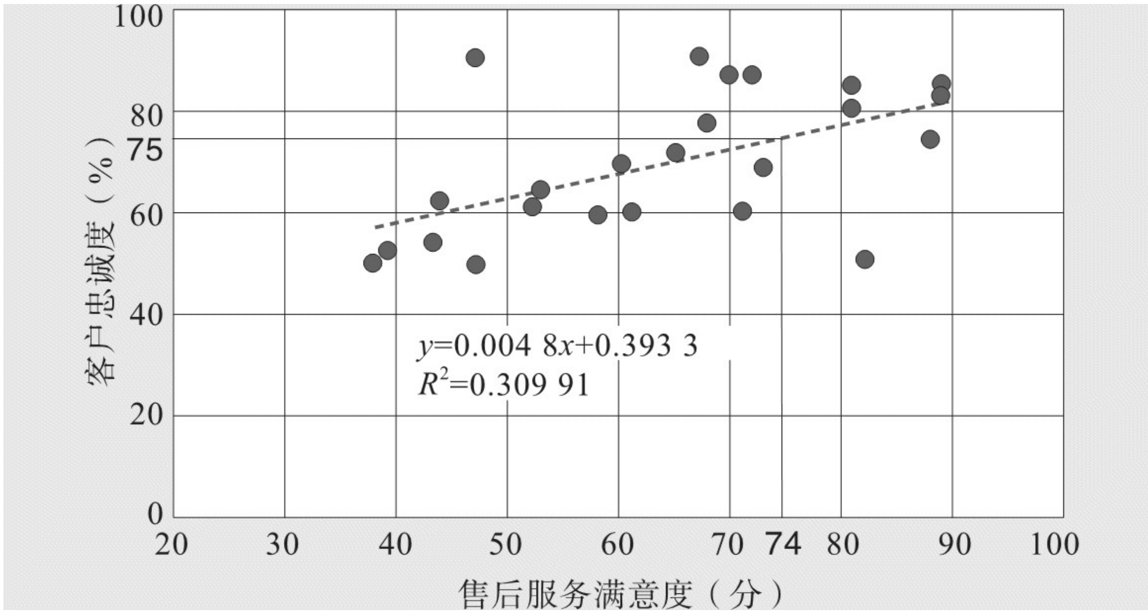
不过这个步骤在车型 B 的部分已经讲过，所以对车型 C 就越过“综合满意度”，直接对“客户忠诚度”与“售后服务满意度”进行回归分析。由此可以计算出售后服务满意度约为 74 分（从图 E 中也可以确认）。

客户忠诚度（75%）=0.0048×售后服务满意度（分）+0.3933

图 D 车型 C 的综合满意度与各方面的相关关系

车型C的客户忠诚度（%）	综合满意度	售后服务满意度	产品满意度	同类产品价格比
83.7	98	89	90	1.29
85.3	100	89	15	1.08
81.1	88	81	25	0.89
91.3	64	67	20	0.93
87.5	82	72	30	1.19
75.0	72	88	90	1.29
87.5	98	70	85	0.92
90.9	68	47	20	1.29
72.4	90	65	100	1.30
85.7	89	81	65	1.15
77.8	91	68	20	0.89
60.6	97	71	55	1.21
61.3	75	52	70	1.12
69.8	65	60	85	1.00
64.9	55	53	85	1.13
69.2	59	73	95	1.02
59.5	69	58	50	0.92
51.2	33	82	90	1.09
61.1	48	61	40	1.05
61.8	48	44	60	1.20
50.0	55	38	80	1.06
54.8	57	43	30	1.08
50.0	44	47	95	1.06
52.6	31	39	15	1.05
与综合满意度的相关系数		0.59	-0.07	0.09

图E客户忠诚度与售后服务满意度的关系



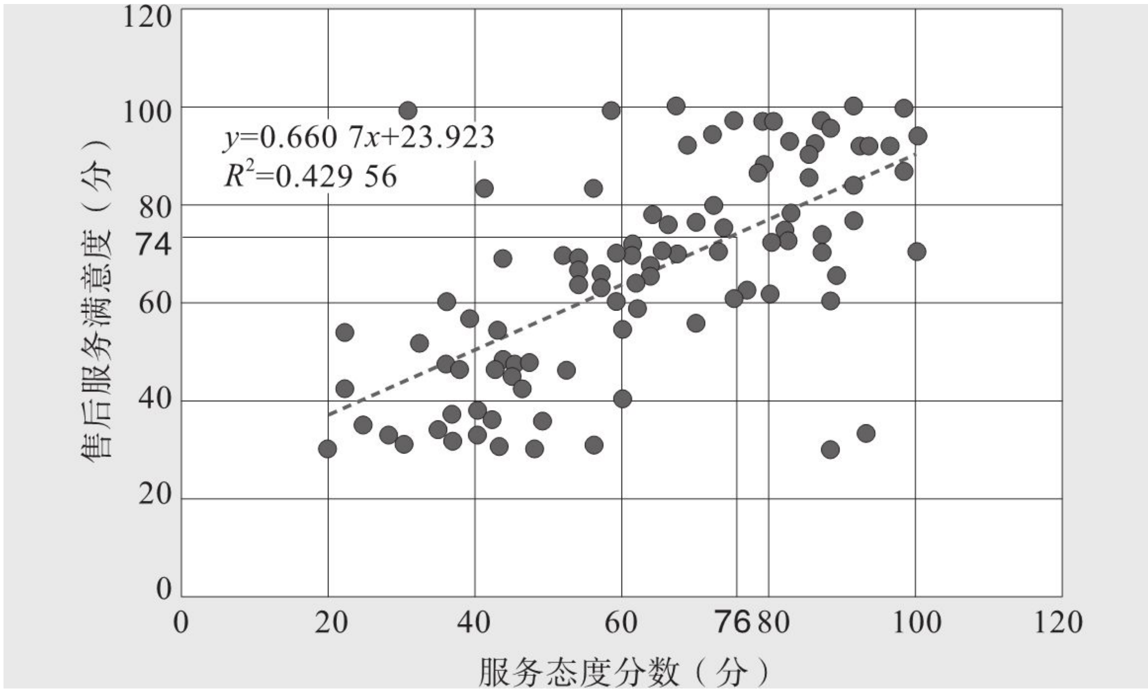
接下来，为了将“售后服务”的概念与更具体的行动措施联系起来，我们调查了售后服务满意度会受到哪些具体因素的影响。图 F 是对 100 名来店顾客进行问卷调查的结果，其中包含了各要素与售后服务满意度的相关系数。

图 F 售后服务满意度与各要素的关联

	电话访问顾客的次数 / 年	服务态度分数	等候时间
	5	44	10
	7	32	14
	5	86	13
	15	87	3
	3	46	3
	15	98	3
	2	54	12
	7	91	13
	7	85	8
	…	…	…
与售后服务满意度的相关系数	0.06	0.66	0.12

从结果可以发现，在 3 个要素当中，接待顾客时“服务态度”的好坏最为关键（相关系数为 0.66）。对售后服务满意度和服务态度进行一元回归分析，其结果如图 G 所示。经过计算可以得知，售后服务满意度要达到 74 分，服务态度分数需要达到 76 分。

图 G 售后服务满意度与服务态度分数的关系（车型 C）

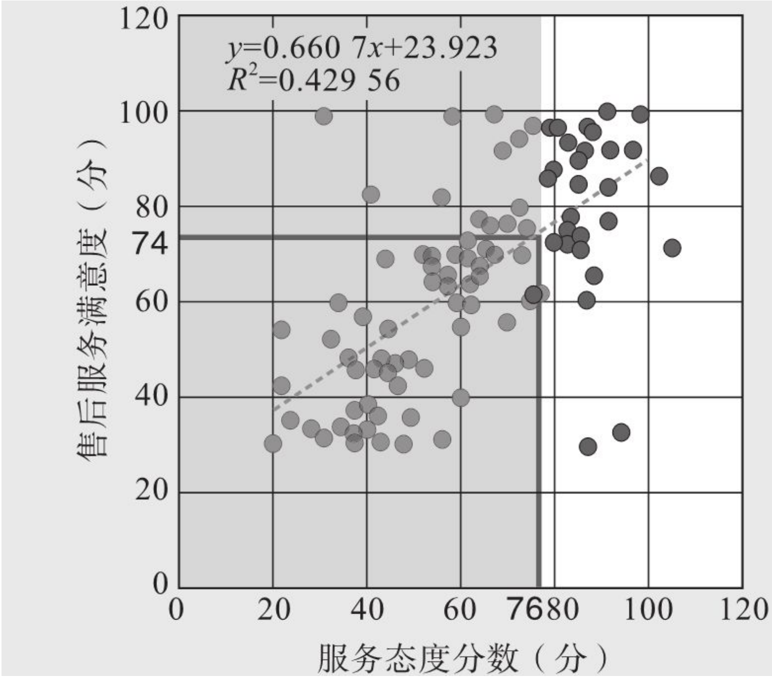


售后服务满意度（74 分）=

0.660 7 ×服务态度分数（分）+ 23.923

那么，接下来需要考虑，采取哪些措施才能将服务态度分数提高至 76 分。首先，从所有服务态度分数当中，将未达到 76 分的分数集中到一起。在 Excel 中使用“IF 函数”，可以筛选出低于 76 分的数值。将分数高于 76 分的单元格清空，计算出其平均值（图 H）。

图 H 售后服务满意度与服务态度分数（76 分以下）的关系



服务态度分数	服务态度分数低于 76 分
44	44
32	32
86	
87	
46	46
98	
54	54
91	
85	
61	61
54	54
39	39
…	…
平均	51

在这个例子中，76 分以下得分的平均值为 51 分。也就是说，我们的目标是将目前低于 76 分的对象的平均值提高 25 分，使之达到 76 分（高于 76 分的分数也会随之提高，所以可以期待整体达到更高的分数）。按照以下条件实施服务培训，以求提高分数的话，我们还可以针对所需资源做出定量预测（计划）。

将 KPI 设定为服务态度分数提高至 76 分（客户忠诚度达到 75%），作为成本，需要规划出 125 万日元的经费（图 I）。

图 I 定量预测实施前提及所需资源

培训的目的及水平	使对象的服务态度得分从平均 51 分提高至 76 分
实施培训水平及预计实施效果	服务态度培训中级（提高 5 分 /1 次）
所需次数	5 次（25 分 ÷ 5 分 /1 次）
单价	1 万日元 / 人次
听讲对象人数	25 人
所需费用总计	125 万日元（1 万日元 × 25 人 × 5 次）

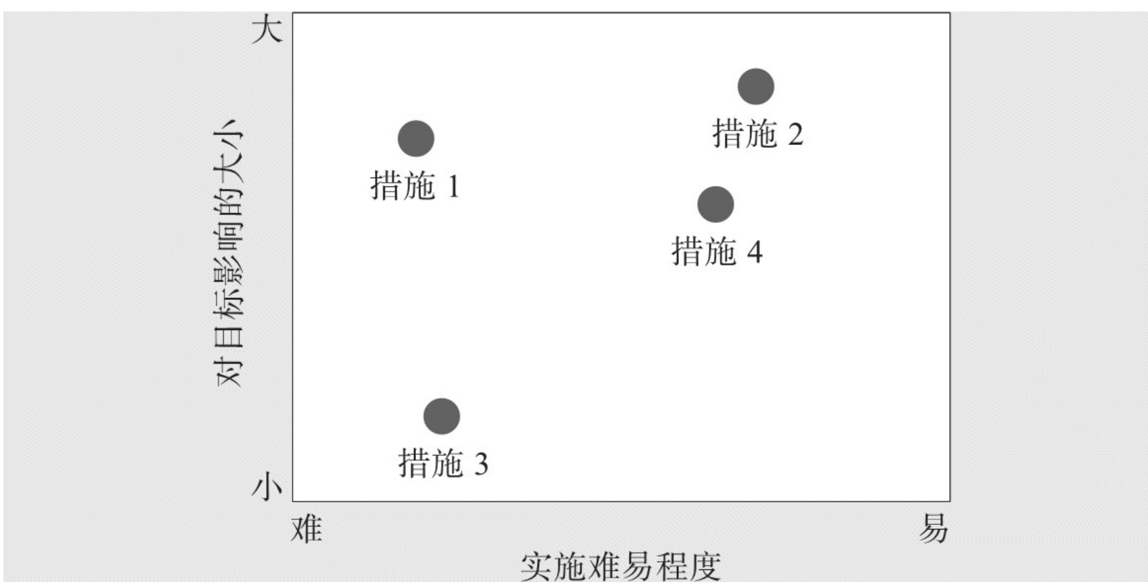
无论实施什么措施，除了设定目标外，在提案中还必须包括费用等所需资源，决策者才能做出判断。否则即使设定了目标，没有最关键的经费来源，也有可能无法实施，而且决策者当然也会问及费用。这种情况下的理想做法也是尽可能以定量的、合乎逻辑的方式，说明为什么需要这些资源（费用、时间、工时等），由此可以实现哪些目标等。

这里只考虑了一个具体的行动措施，不过很多情况下需要讨论多种措施。这时就需要列出优先顺序，明确指出有限的资源能够实现的范围和实施顺序。决定优先顺序的标准因具体情况而异，一般可以按以下维度简单列举出来（图 J）。

•对目标影响的大小

•实施难易程度（所需时间、技能、工时、总费用等较少的情况更为理想）

图 J 确定优先顺序的矩阵图 1



假设各措施可以形成图 J 所示的位置关系。位于右上方的措施最为理想，即“最易于实施，效果（影响）最大”。同样，将斜线逐渐向左下方移动，可以确定接下来的优先顺序。因此，前面列举的 4 个措施可以确定以下优先顺序（图 K、L）。

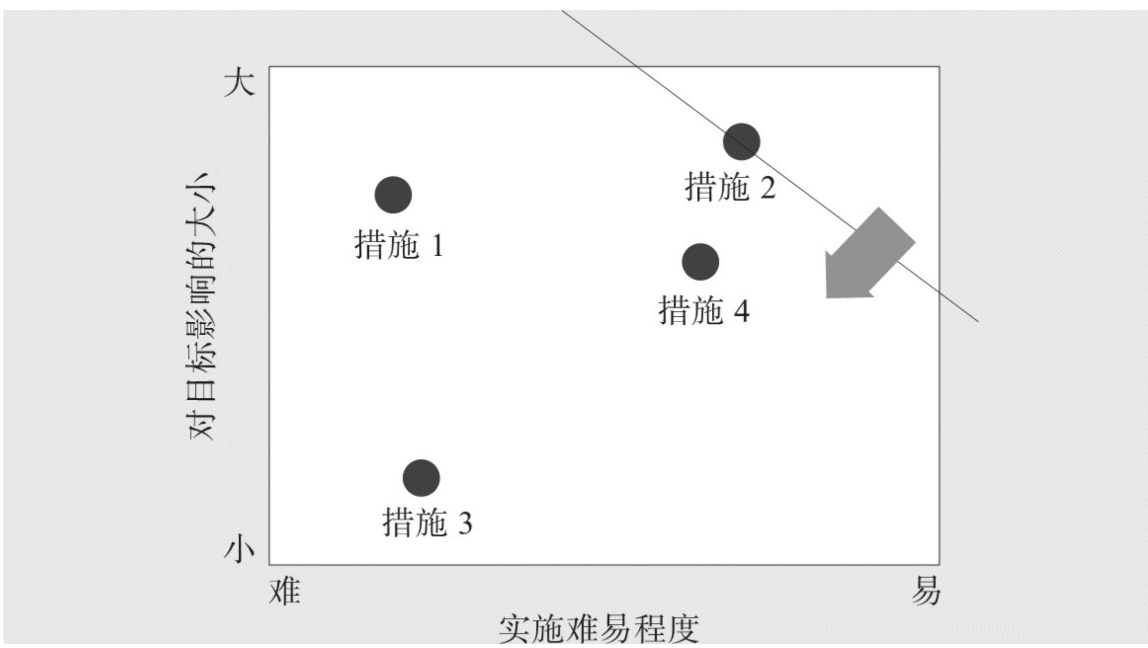
优先顺序 1：措施 2

优先顺序 2：措施 4

优先顺序 3：措施 1

优先顺序 4：措施 3

图 K 确定优先顺序的矩阵图 2

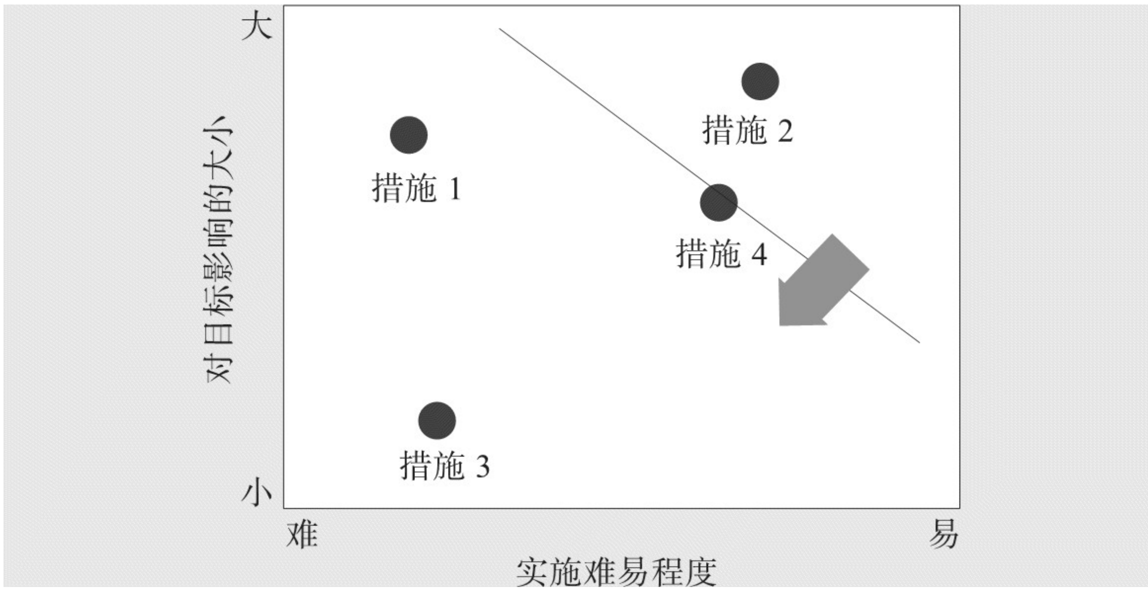


在这个过程中，如果资源有限或存在其他方面的制约，便可以首先集中力量实施措施 2 和措施 4。

各项措施的定位不一定要非常严密，不过为了在实施阶段得到相关人员的认可，最好是经过多人确认，最后落实为大家一致同意

的内容。此外，运用这种图表（支付矩阵），从直观上展现优先顺序的决定过程，也更容易得到决策者的理解，可谓一举两得。

图 L 确定优先顺序的矩阵图 3



接下来的第 5 章会回顾之前的一系列“解决问题的故事”，作为对本书的总结。

第 5 章
用数据讲故事

根据每章篇末的“解决问题的故事”，不知道大家有没有将每个单独的分析过程串联起来，形成一个完整的解决问题的故事？最后，我们再将前面的要点（故事），分为 4 个部分回顾一遍。

（1）第 1 章和第 2 章的重点是“明确目的或问题”及“大致把握现状”（图 5-1）。与去年一年的月平均销售（额）相比，本年度最近一个月的新车销售（额）减少了约 15%。对比本年度计划，可以确认约有 10% 的目标无法实现（图 5-2）。

图 5-1 明确目的或问题以及大致把握现状

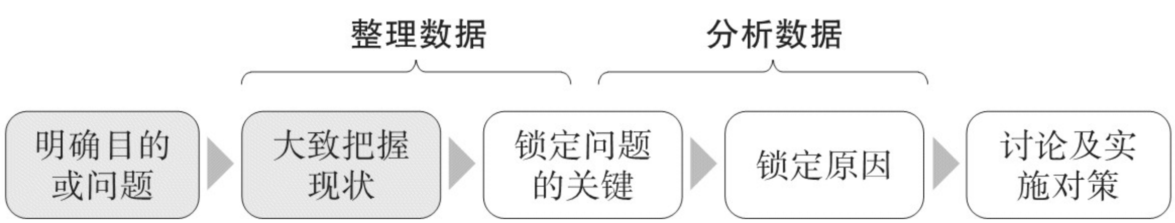
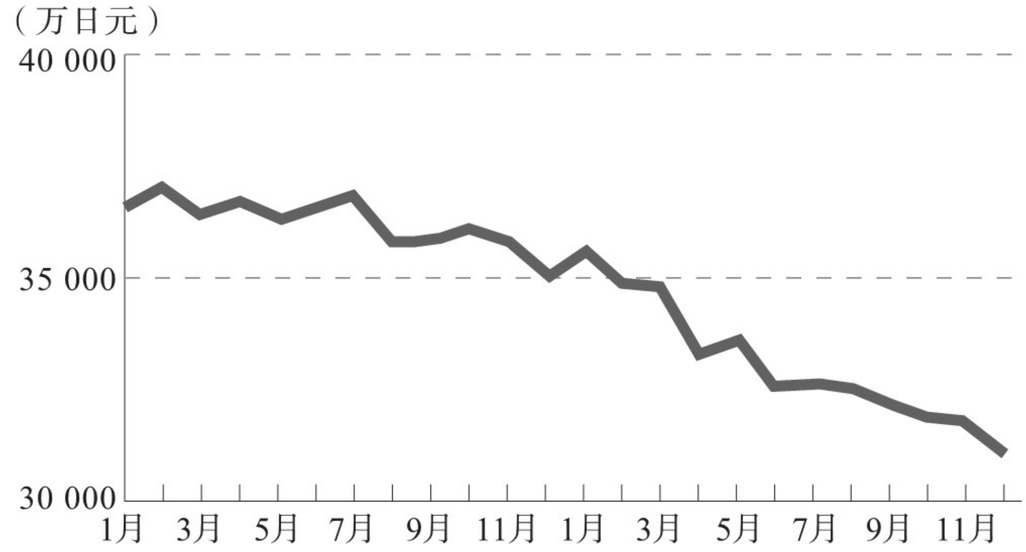


图 5-2 新车销售总额的变化



（2）第 2 章的重点是提出具体的论点，锁定问题的关键（图 5-3）。得出“问题在于客户忠诚度”的结论之前，我们经过了以下讨论：

•新车销售额减少不是价格（保持在±5% 之间）降低造成的，而是销售数量减少造成的（图 5-4）。

•销售数量减少不是首次购买数量（基本维持不变）的减少造成的，而是再次购买数量的减少造成的（图 5-5）。

•再次购买数量的减少不是再次购买对象人数的减少造成的，而是对本品牌产品的再次购买率（客户忠诚度）降低造成的。也就是说，顾客在再次购买汽车时，选择其他公司产品的比例增高了。

图 5-3 锁定问题的关键



图 5-4 新车销售数量的变化

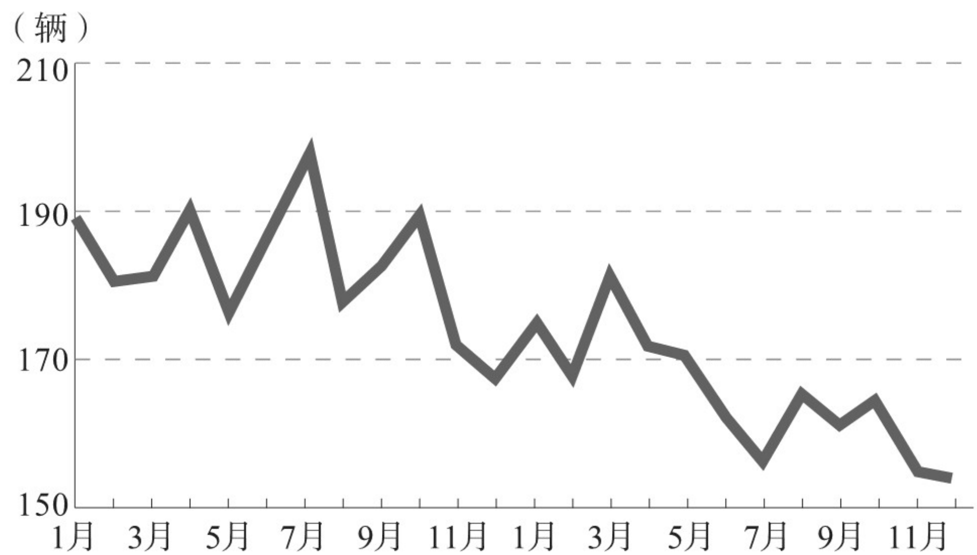
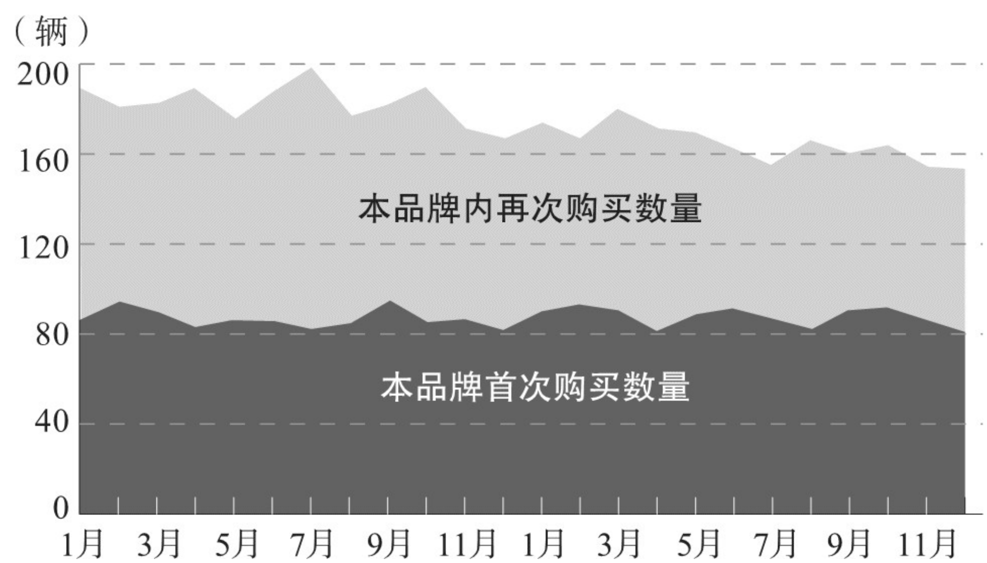


图 5-5 新车销售数量明细的变化



接下来，将问题关键锁定为“车型 B 及车型 C 的客户忠诚度”的过程如下：

•根据对整体销售数量的影响，应该关注的对象车型依次为车型 B、车型 C 和车型 D（图 5-6）。

•其中，客户忠诚度下降的是车型 B 和车型 C（图 5-7）。

图 5-6 各车型新车销售比例

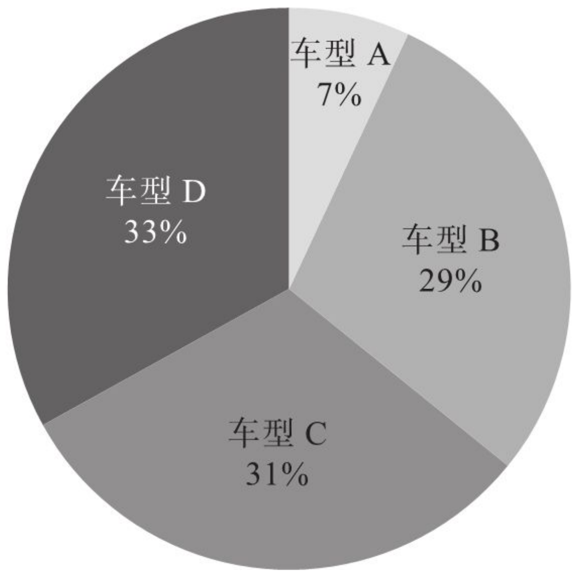
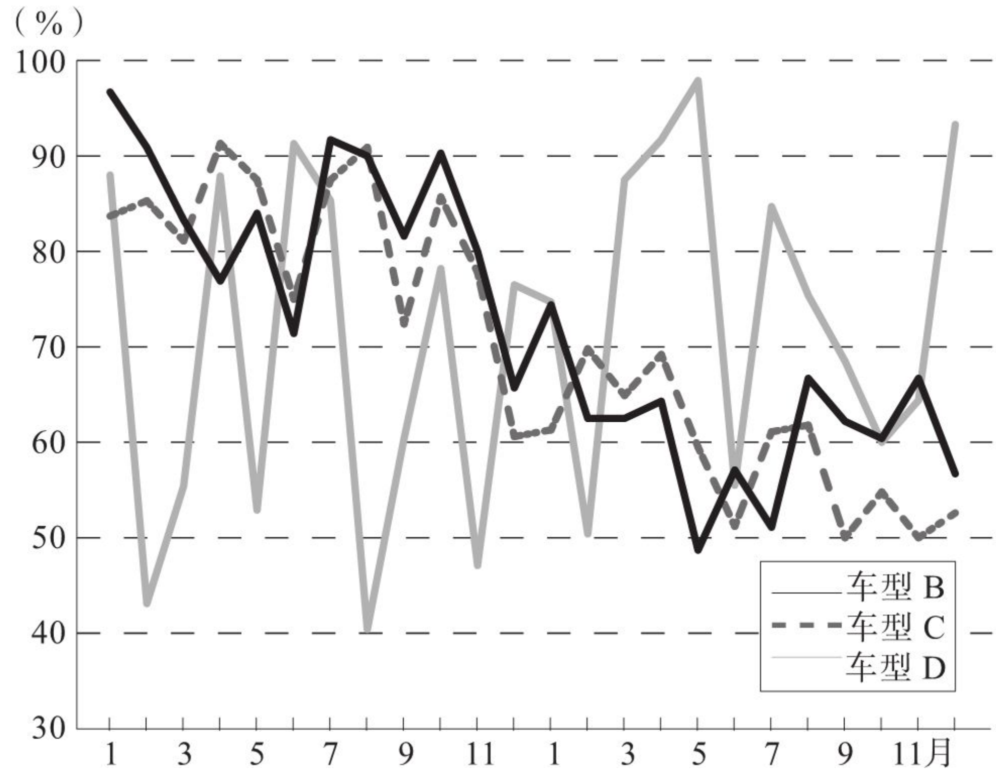
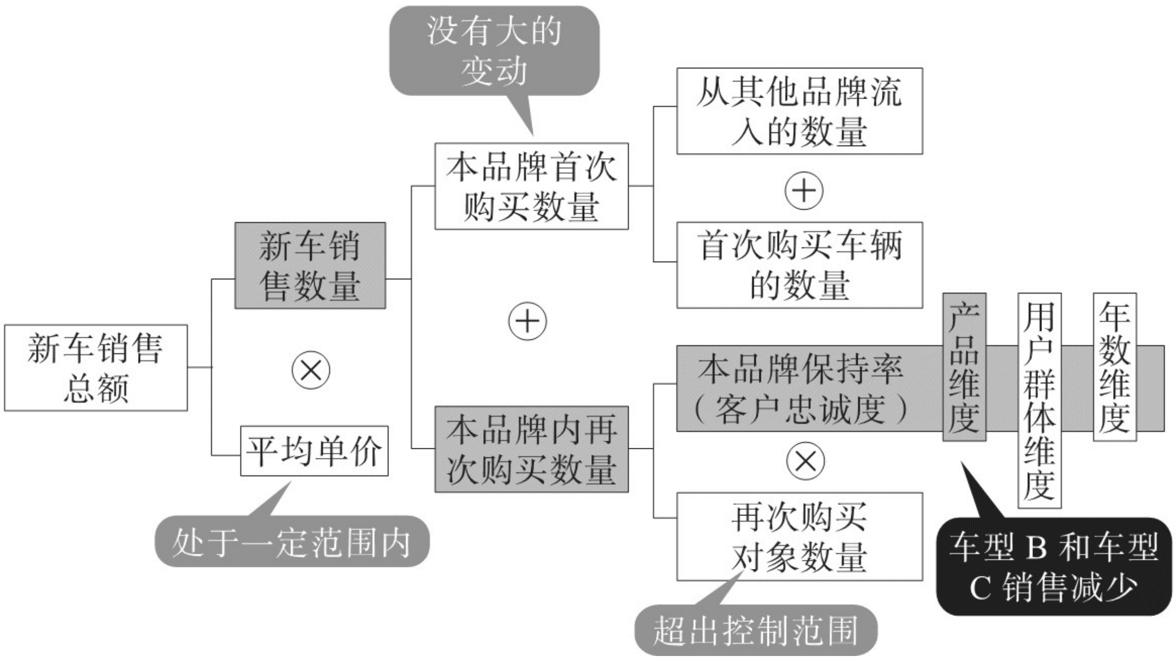


图 5-7 各车型客户忠诚度的变化



根据上述内容，我们将问题关键锁定为“车型 B 和车型 C 的客户忠诚度”。至此为止的整体逻辑结构如图 5-8 所示。

图 5-8 锁定问题关键的逻辑构造



（3）接下来，第 3 章的重点为如何“锁定原因”（图 5-9）。在这个过程中发现了以下三点内容。

•通过相关分析，得知客户忠诚度与顾客综合满意度相关

•关于车型 B，得知“同类产品价格比”对“综合满意度” 影响最大

•关于车型 C，得知“售后服务满意度”对“综合满意度” 影响最大

也就是说，通过以上过程，发现销售减少的主要原因，车型 B 是“与竞争对手的价格比”，车型 C 是“售后服务满意度”。经过进一步挖掘，还可以得知车型 C 的“售后服务满意度”受“服务态度的”的影响最大（图 5-10）。

图 5-9 锁定原因

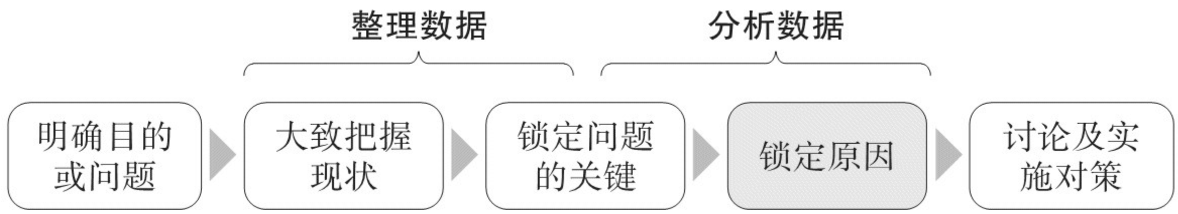
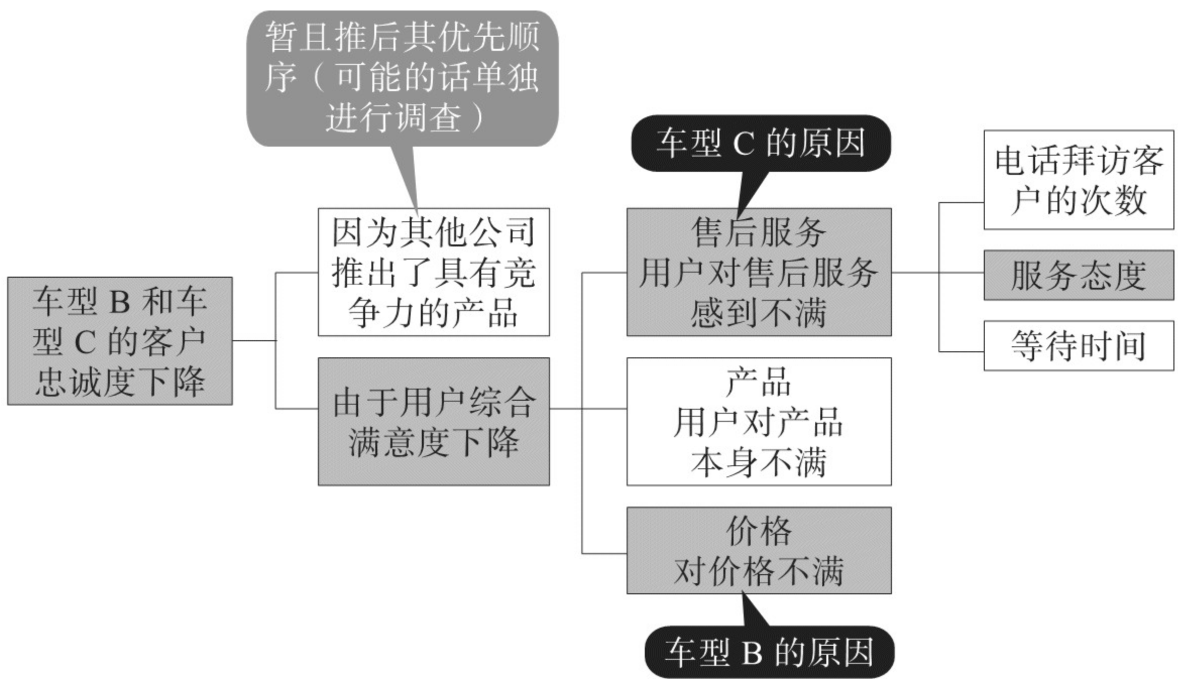


图 5-10 深入挖掘客户忠诚度下降的原因



此外，还有一些要点未能在上述分析中涉及，如“再次购买对象数量”的变化，其他公司产品可能对客户忠诚度产生的影响等。

针对这些内容，本来还应该做进一步调查。因为这些都是公司外部的原因，往往容易被漏掉。与外界信息相比，公司内部的数据当然更容易获得，也更容易受到关注。人们的意识常常只集中在能够看到的范围之内。我们在随时关注身边信息的同时，还应该养成习惯，经常提醒自己“还应该关注其他范围”。

（4）第 4 章是最后“讨论及实施对策”的阶段。根据回归分析的结果，车型 B 为了使客户忠诚度达到 75%，设定了以下目标值及行动措

施，并传达给各位销售人员（图 5-11）。

KPI：使综合满意度提高至 75 分

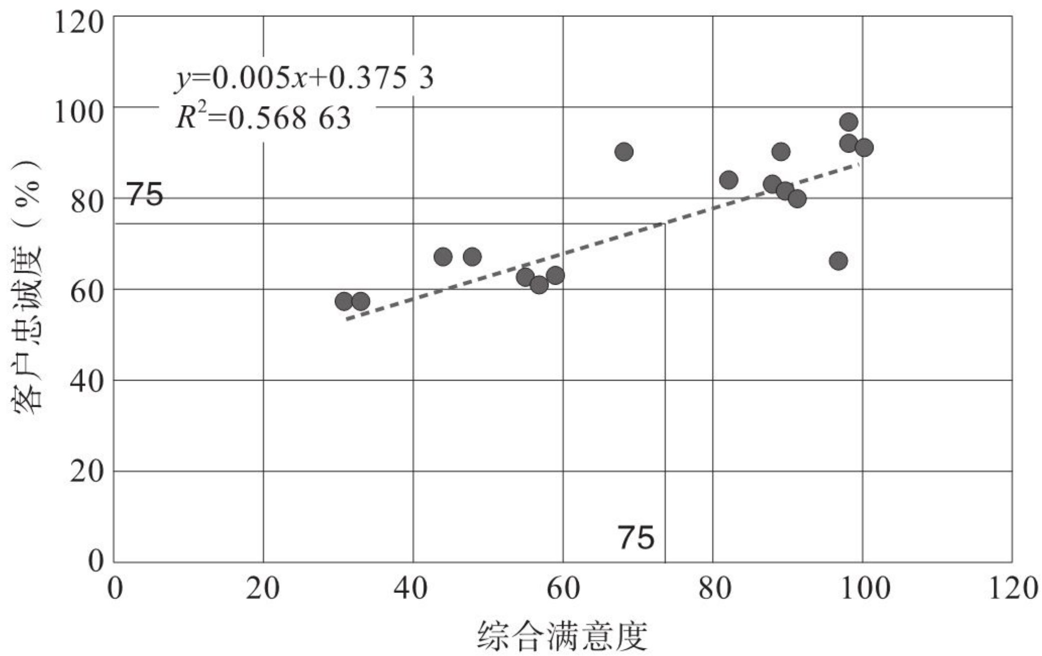
行动措施：将同类产品价格比控制在+2% 以内

同样，对于车型 C，也设定了以下目标值及措施（图 5-12）。同时还计算出，需要确保有 125 万日元的预算作为培训费用（以 25 人为对象，实施 5 次）。

KPI：使售后服务满意度达到 74 分

行动措施：实施培训，将服务态度分数提高到 76 分以上

图 5-11 为实现目标设定 KPI（车型 B）



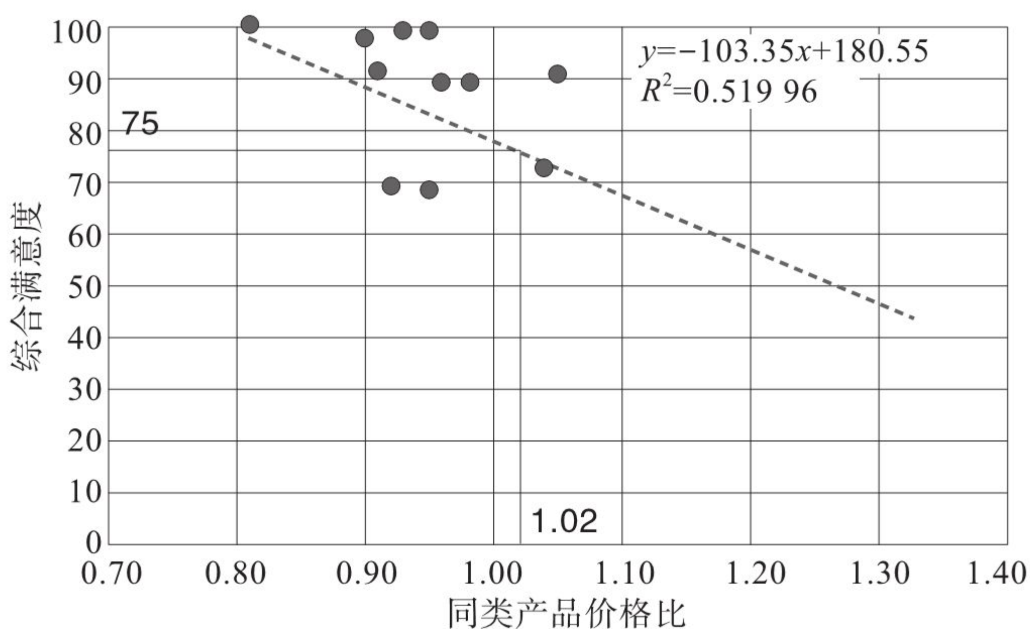
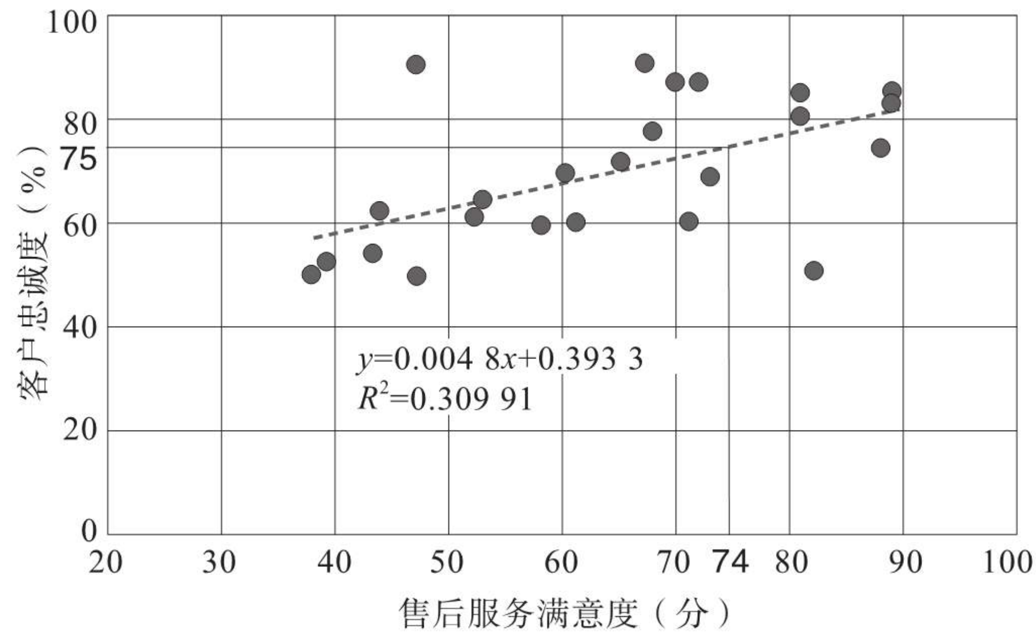
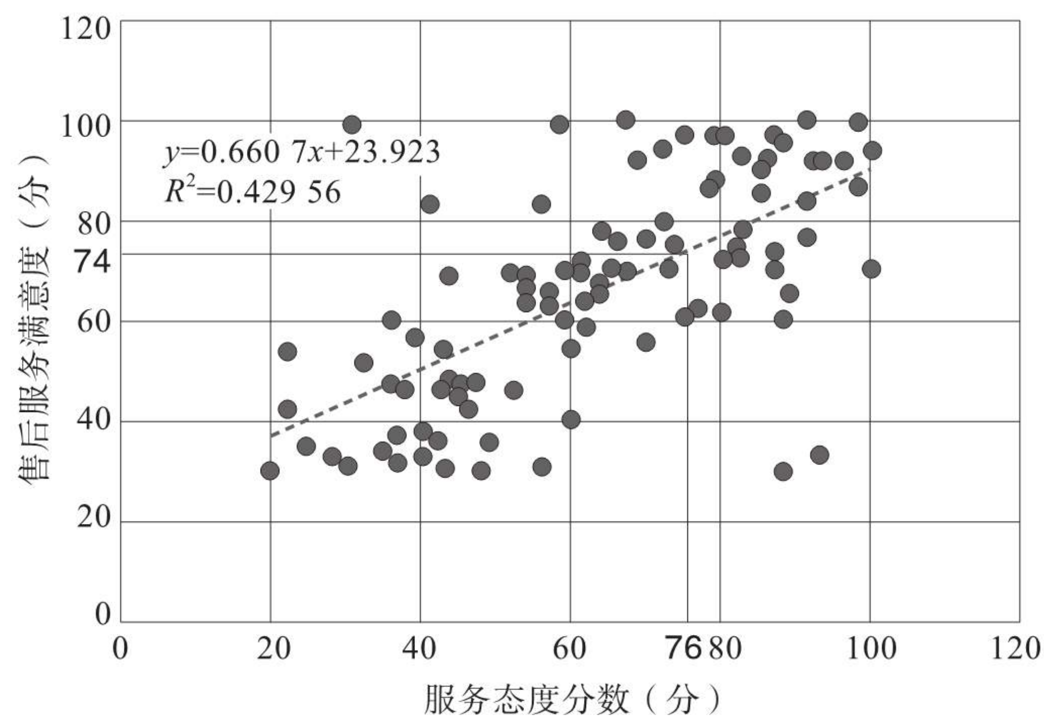


图 5-12 为实现目标设定 KPI（车型 C）



100
80
75
60
40
20
0
客户忠诚度（%）
y=0.004 8x+0.393 3
R²=0.309 91
20
30
40
50
60
70
74
80
90
100
售后服务满意度（分）



120
100
80
74
60
40
20
0
售后服务满意度（分）
y=0.660 7x+23.923
R²=0.429 56
0
20
40
60
76
80
100
120
服务态度分数（分）

以上解决问题的故事可以总结为“明确目的或问题”→“大致把握现状”→“锁定问题的关键”→“锁定原因”→“设定行动措施及 KPI（和所需资源）”。其实，这与医生为患者看病并开出处方的过程有很多共同点（图 5-13）。

图 5-13 医生看病与解决问题的过程的比较

过　程	提出问题	医生看病	解决问题的过程
明确目的或问题，大致把握现状	表现为哪些现象（症状）	自前天起腹痛	新车销售额比前年减少了 15%
锁定问题的关键	哪里出了问题	胃痛	问题为车型 B 和车型 C 的客户忠诚度
锁定原因	为什么会出现这个问题	吃得过多	车型 B：同类产品价格比的问题 车型 C：服务态度的问题
研究及实施对策	需要何种对策	一周内控制饭量，服用胃药	车型 B：设定价格比上限 车型 C：进行服务态度培训

过　程	提出问题	医生看病	解决问题的过程
明确目的或问题，大致把握现状	表现为哪些现象（症状）	自前天起腹痛	新车销售额比前年减少了 15%
锁定问题的关键	哪里出了问题	胃痛	问题为车型 B 和车型 C 的客户忠诚度
锁定原因	为什么会出现这个问题	吃得过多	车型 B：同类产品价格比的问题 车型 C：服务态度的问题
研究及实施对策	需要何种对策	一周内控制饭量，服用胃药	车型 B：设定价格比上限 车型 C：进行服务态度培训

无论解决什么问题，都需要这样一步步锁定对象，摸索原因。不过实际工作中，我们未必能马上获得自己所需的数据，常需要对假设进行重新调整，反复尝试多个回合。在假设和检验的过程中循序渐

进，将所发现的事实和逻辑一层层积累起来，这个过程对任何问题来说都是相同的。

那么，这些工作对解决问题能产生多大作用呢？数据分析并不是到考虑对策就结束了。本书没有针对所有问题讨论所有的对策，不过我们可以粗略地了解一下，前面 4 章的内容能够对解决问题产生多少效果。

首先，最近的销售额减少到了 3.1 亿日元。其中首次购买部分与再次购买部分各占一半，所以假设再次购买部分为 50%。根据各车型销售比例饼状图可以得知，车型 B 与车型 C 加起来共占 60%。本次主要针对这 60% 的部分锁定原因，并提出了改善措施。假设通过一系列措施实现了公司决策层设定的目标，成功地将客户忠诚度提高了 25%，则可以按照下面的公式计算出由此增加的销售额。

在细节方面，计算比例时也许可以采用不同的计算方法（以何种数据为母体、以金额还是以数量计算等），也可能会有误差。不过只要能把握大致规模，基本上这样就可以了。

31 000（万日元）×50%（再次购买部分）×60%（车型 B 与车型 C 所占比例）×25%（通过相应措施得到改善的部分）=2 325（万日元）

这意味着公司的销售增加了 7.5%（= 2 325÷31 000）。最初的问题是销售额减少了 15%，那么现在这个结果就是可以恢复一半。以现有资源为前提，使销售额的减少额恢复一半，这样的成绩还是相当不错的。接下来，下一期可以在此基础上完成剩余的一半（或者同时采取其他措施，完成剩余的一半）。

在解决问题的过程中，上级目标被落实为具体的 KPI，既便于业务一线人员理解和运用这些 KPI，同时也有利于管理者对最终影响进行定量把握。

如果不知道所采取的措施能产生多大作用就贸然实施，目的就有可能变成了“实施”。既然是从定量的角度解决问题，当然最好也能从定量的角度跟进所采取的措施和结果。

把解决问题的过程展现出来

需要将措施或预算提交给决策者时，将相关信息条理清晰地展现出来，会取得更好的效果。当然，对于有些听众来说，先陈述结论（希望得到批准的措施、预算等）会比先陈述理由（锁定的原因）效果更好一些。

我在日产工作时，会根据对象是外国人还是日本人，而有意改变结论的位置。日本人大多习惯于从头开始依次说明，把想说的结论放在最后。但对外国人照搬这套流程，则很可能会引起对方的烦躁。可能还不到几分钟，现场就会陷入“你到底想说什么”的气氛中。

既然目标是让对方听了自己的说明，能够理解内容并做出决定，那么就应该根据对方的情况在表达方式上“下点功夫”，这一点在多元化组织中尤为重要。不过在任何情形下，都不要打乱得出结论的逻辑和流程（即故事）。

此外，与解决问题的中间过程、分析内容相比，对方可能更想先知道结论及其根据。因此，面面俱到地介绍分析过程中的所有详细内容未必是上策。多数情况下，这些内容都是在被问到的时候稍加说明就足够了。

有些人制作演示资料或者报告时，会密密麻麻地写满详细的文字或图表。这种“不知道哪些地方会被问到，所以就全都写上”的心态，还是放弃为好。我们应该把时间和精力用来考虑“核心信息是什么”“对方想知道什么”。

在组织中使用数据的价值与难点

越来越多的人认识到，当今时代已经不能再依靠以人为本的“老方法”取得成果了。新的环境要求人们采用全新的工作方法。人们不能再凭直觉或经验，随随便便地决定工作怎么做了（过去也许可以）。

世界的复杂程度和前进速度都今非昔比，仅靠自己或所属组织的经验已经无法应对所有问题。值得庆幸的是，收集数据的技术和系统也获得了长足发展，我们可以充分利用这些有利的条件。

做到这一点，并不一定需要高难度的分析方法或者昂贵的系统投资。提高工作当中的数据运用能力，即提高员工的分析技能，从长远来看也有助于增强公司的竞争力。

日产公司长年坚持在世界各地的组织中推行提高“解决问题能力”的活动。因此逐渐形成了一种机制，确保不同工作岗位、不同国家之间拥有共同的流程或语言。经过共同的流程得出的结论，不用经过决策者逐一确认，也能在一定程度上保证其可信度及内容的水平。

不过尽管日产如此积极地推行“使用数据”的工作方法，除了部分专业部门之外，大部分员工在这方面的能力仍然远远不够。外籍经理或高层管理者的要求不是“循序渐进地学会使用数据”，而是“从明天开始就要在提案中体现出来”。“用数字工作”已经与“用英语交流”一样，成为员工必须具备的条件。

为了确保多元化人才聚集的大型组织能够顺利决策，迅速有效地推动业务发展，数据分析已经被视为一项核心技能。

无论是未知领域的业务，还是新组建的事业，只靠自己过去的经验都未必能得出正确的答案。个人的经验或知识实在有限，仅凭这些已经无法确保工作顺利进展。

通过有效运用外界的数据信息，无论是个人还是组织都能够获得爆炸式增长。同时，及时发现效率低下或徒劳无功的工作，不断加以

改进，还可以削减那些“本来已经没有效果，只是为了维持现状而运行”的项目等。无论是对希望用更少的人员获得更多业绩的公司来说，还是对在公司中工作的个人来说，这都将是一个幸福的结局。

不过，也有一些人或组织很难从心理上立刻接受这种用符合逻辑的客观数据来推进工作的方法。有时甚至还会出现“靠纸上的数字能知道什么”“只说些冠冕堂皇的漂亮话，什么问题也解决不了”等否定态度。这种情形下，即使努力进行分析，仍然有可能会碰壁。

尤其是到了“解决问题”的阶段，不可避免地要得出“问题在于XX”“是 XX 的问题”等结论。这时，很难想象相当于 XX 的部门或负责人会心悦诚服地接受这个结论。

我在专门负责日产内部改革的部门工作时，高层管理者要求我推行彻底的“改革”，而不是改善。改善说到底仍处于现状的延长线上，只是变得更好，因此人们的抵触也比较小。而改革大多是要否定现状，创造出新的东西，对组织和个人的影响都比较大。这种情况下，如果直接用数字抛出符合逻辑的道理，几乎不可能取得预期效果。

因此，我会尽量避免一个人依照本书介绍的这一系列流程（虽然从理论上、技术上我可以做到）开展工作。我总是随时与相关人员分享分析的过程，必要时与他们展开讨论，了解对方的想法和困惑。此外，我还注意花时间把自己的工作详细介绍给对方。因为如果等到一切“答案”都揭晓以后，才突然把结论告诉对方的话，可能会遇到远远超出想象的抵触。

与问题相关的人未必都持有合理且客观的目标。与公司或组织的利益相比，他最优先考虑的可能是自身利益。所以我也需要了解关键人物的真正的目标是什么，包括公司内部的政治背景，从而尽量在他的目标与解决问题的合理目标之间找到共同点。

当然，并不是所有的情况下都能找到合适的解决方法。有时双方的分歧无法消除，只能到最后关头由最高决策者来判断。这时，依据

客观数据把握现状并锁定原因才更具有绝对价值。因为双方感情用事相互顶撞，是无法做出合理判断的。

这些因素不只是专业的数据分析人员，更是“商务人士”所特有的技能，是“干好工作”的条件之一。

你能用数字推翻众人的理解吗

我在日产工作期间，遇到过最困难的工作之一是“用数字来证明多数人的理解都是错的”。我当时的工作是针对高管认定的问题进行分析并提出对策。这种情况下，可以收集一些相关数据，做做表面文章，然后归纳出一个故事，汇报说“确实是这里有问题，因此我认为应该这样来改进”等。这样做最“轻松”，也不会在公司里掀起波澜，是一种“稳妥”的方法。

但我依据自己的亲身体验，实在无法认为高管指出的就是真正的“问题”。于是我收集了大量数据，最后证明“问题不在这里”，得出了完全相反的结论。“XX 不是 XX” 就好比“没有鬼”一样，是很难证明的。

此外，组织内部还有一些压力，例如认为“明明是高管指出的问题，你（提出反对意见）是作何打算”等。因此无论从技术上，还是从心理上，我都要面对一个棘手的难题。

我并没有这个问题所涉及领域的实际业务经验，不过这次经历教会我的是，只要认真收集数据，条理清晰地解读事实，就算无法 100% 证明正确，也仍然能够得出具有一定可信度的结论。对于大家都在主观上认为是“A”的问题，我凭借数据的力量，提出“不是 A，而是 B”的结论。

当时有一位外籍董事支持我的观点，他主动替我向最高层主管进行了演示和说明。就算是为了不辜负他为我承担的风险，我也要仔细地确认事实，构建出严密的逻辑根据，以求万无一失。这项提案果然获得了高层的认可。

由于我的坚持和努力，大家没有成为“本来没有问题的问题”的牺牲。此外，因为需要各个方面的数据、信息以及协助，我通过这个项目得以在公司构建起牢固的人脉网络。一年半以后，我没有任何经验却获得了财务部的职位，这是我在日产工作的最后一个岗位，也是当时那位外籍董事介绍给我的。

更上一层楼（高级技能简介）

前面介绍的这些技能，能够用来解决所有行业中的问题。那么掌握了这些技能之后，还可以再学习、积累哪些内容呢?

接下来我们既可以力争解决多种问题，积累经验，也可以掌握一些难度更高的技能。作为更高一级的内容，我想介绍 2 种方法。感兴趣的读者也可以阅读相关书籍来加深理解。

（1）多元回归分析

第 4 章介绍的回归分析因为只涉及 2 种数据，所以被称作“一元”回归分析。一元回归分析可以说是将现实世界极度简化的一种模型。对有些分析所要求的准确度或目的来说，这种模型就足够了。

不过，例如考虑“销售额”时，恐怕无法只用产品、价格、顾客等当中的 1 个原因进行说明，而是需要分析多个原因的综合作用，这样才更接近现实。

“多元”回归分析的方法可以实现这个目标。与简单的一元回归分析相比，多元回归分析能够考虑更为复杂的要素，难度更高，其结果

也更具有说服力。我们可以将一元回归分析和多元回归分析得到的结果放在一起，对比其不同之处（分析结果为虚拟数值）。

一元回归分析：销售额（日元）=来店人数（人）×35.2+50 000

多元回归分析：销售额（日元）=来店人数（人）×40.8 +减价率（%）×14.9 +宣传单投放频率（次/周）× 20.5 + 85 460

看到这里，可能有人会说，“那么就不要用一元回归分析，都用多元回归分析吧”。然而考虑到实际情况，并不能这样说。就我自身而言，除了极为特殊的情况，一般不会在培训中介绍多元回归分析，或者在实际工作中推荐别人使用多元回归分析。

在 Excel 加载项中有“分析工具”，其中包括回归分析，我们可以用它来进行分析（详细请参照本书 105 页对 Excel 加载项的介绍）。试用一下就可以发现，多元回归分析的结果恐怕会让很多人都大吃一惊，而我也完全能理解他们的心情。

因为其结果并不是用“销售额=……”的公式来表示的，而是一下子列出“自由度”“必然 F”“标准误差”“t”“p 值”等很多让初学者不知所措的专业词汇。这些数值主要是用来评价分析结果是否可靠的。

如果只看“系数”栏，也可以写成方程式，但其前提是要对该方程式在统计学上的可信度进行评价，之后才能用于实际工作。要想得出准确的分析结果，这一步是不能省的。

具体方法是看各系数的“p 值”，当它低于 5%（0.05）时，该系数为零（即该系数不产生影响）的可能性在统计上很小，因此可以作为回归分析的结果来使用。相反，如果“p 值”超过 5%，则需要剔除这个系数，重新进行回归分析。

为什么需要考虑 p 值呢？其主要原因与“多重共线性”有关。多重共线性（multicollinearity）指如果多元回归分析的解释变量之间存在相关关系，其相互影响会导致分析得到的系数失真。

例如前文对“销售额”的分析中，如果“宣传单的投入频率”与“来店人数”各自独立的话没有问题，但如果投入宣传单也会对来店人数产生影响的话，就有可能引起多重共线性的问题。也就是说，多元回归分析的前提是，解释变量之间不相关，各自独立。否则，就无法准确计算出各变量的系数。

按理来说，我们可以先计算解释变量之间的相关关系，然后只选择不相关或相关较低的变量进行多元回归分析。但在实际工作中，这个问题非常棘手。

为了确保不产生多重共线性问题，需要反复改变变量的组合，多次进行多元回归分析，确认整体结果，这样会耗费大量精力。而且如果数据有限，那么也很可能出现下面的情况，即尝试了现有数据的各种组合，但都存在多重共线性问题，数据根本派不上用场。

我也曾经有幸找到合适的变量组合，从而得出高准确度和可信度的分析结果，但也有过很多次大费周章最后却徒劳无功的经历。

对多重共线性、分析结果的统计学评价等问题，如果分析者不具备一定程度的知识和经验，在遇到“这个结果真的正确吗”的质疑时，就只能含糊其辞地回答“应该是的……”。在事关重要经营判断的场合，一知半解地使用多元回归分析得出结果，从各个角度来看都伴随着较大风险。

虽然写了这些批判性的观点，但我并非完全反对使用多元回归分析。在理解了这些难点和前提条件的基础上，运用难度更高的数据分析，有时也可以获得更多的知识。

此外，多元回归分析是多变量分析的典型方法。不用自己分析，而是将分析工作外包给外部公司去做时，为了便于与专业人士沟通，或者更好地理解收到的分析结果，我们都可以了解一些多元回归分析的方法和注意事项。

下面再介绍一下多元回归分析广阔的应用范围及其真正实力。多元回归分析也分几种不同类型。例如数量化一类分析可以定量表示男

女、天气、季节（春、夏、秋、冬）等定性变量，对其进行回归分析。该方法将定性项目置换为虚拟变量“1”和“0”。例如用 1 表示男性，用 0 表示女性。通过回归分析，可以得出类似“下雨天销售额会增加 10 万日元”“春季来店人数会减少 500 人”等结论。

此外，在市场营销领域，还有叫作“logistic 回归分析”的方法。该方法可以用于被解释变量是定性变量的情形。例如预测顾客对某个活动做出反应的概率等。这种分析所需的数学背景要比基于比例关系的普通回归分析更为复杂。

除了 Excel，还可利用各种应用软件进行多元回归分析。但即便使用应用软件，也需要分析者对包括多重共线性问题等分析做出必要的判断和调整。我们不能把数据导入软件，就毫无保留地信任和使用软件得出的答案。为此，也需要具备基本的数据理解能力和分析能力。

（2）样本与总体（假设检验）

我们要明白，用于分析的数据并不代表“全部”，这一点极其重要。例如，假设我们想使用会员数据来分析顾客的购买行为。但这些数据只是愿意成为会员的“积极顾客”。即使能够收集到所有顾客的数据，也仍然不能确定它对于了解新顾客的特性可以发挥何种程度的作用。

这里需要注意两件事。一是在实际工作中，用于分析的数据大多是“现有数据”。因为是能够获得的有限数据，所以就有可能会产生偏颇。还有一个注意事项，即使是根据分析目的，有计划地调查或收集到的数据，它们也都是样本数据，而不是所有数据。

这些理所当然的事情，却常常被忽视。无论使用多么高级的分析工具，进行多么无懈可击的分析，如果最关键的输入数据有问题，其结果就会质量堪忧。

对于上述第一个注意事项，并没有根本的解决方法，因为“没有的东西（数据）就是没有”，不过至少分析者需要了解，自己所采用的数据属于哪个范围，它对于目的来说处于何种定位。很多情况下，我们

获得的数据大多只是与“眼前的问题”相关的数据，例如“新宿店的商品A的销售数据”等。可能我们想针对商品A或者新宿店销售减少的状况进行分析，但仅凭这项数据只能获得半径几十米范围之内的信息。

如果能获取其他店铺、其他商品等数据，分析的视点就可以扩大到“公司内部”。接下来应该考虑的是，“只是本公司内部的问题吗”，这样视点就能扩展到与外部的竞争。于是还需要竞争对手、同类产品的数据。

接下来，如果要考虑开发过去对这种商品及竞争对手产品都不感兴趣的新顾客，从而改善销售情况的话，只看销售业绩的数据还不够。因为销售数据显示的只是已经购买该商品的顾客的行为和判断。这个范围之外的人也许会做出其他判断或行为。此时就需要另行调查或者收集其他数据。

那么，关于样本数据应该怎么做呢？我们需要了解统计分为“描述性统计”和“推断性统计”。描述统计指使用平均值、标准差等统计指标来整理或展现多个数据的特征。推断统计指根据作为样本收集的数据，来推断所有数据（总体）的特征。

大多数分析所用的数据都是样本数据，因此必须了解分析的前提是用样本数据来代表总体。严格来说，采集数据样本时也需要注意一些问题，防止出现偏差。

不过在实际的日常工作中，常常由于时间和精力有限，即使在一定程度上牺牲准确度，也会利用现有的数据进行分析。也就是说，学术上的严谨性与实际工作之间需要达到平衡。

虽然大数据时代有可能为我们提供庞大的数据用于分析，甚至不用区分样本与总体，但从一般的工作角度来看，可以说这只是特殊情况。因此，作为根据样本数据推断总体特性的方法，接下来介绍被称为“检验”的统计方法。

例如，假设让随机抽选的男女顾客各50人为新开发的糖果产品打分，对平均分进行比较。结果是女性顾客的平均分高于男性。那么是

否可以认为“该产品更受女性欢迎”，把重点放在面向女性的广告上呢？如果不考虑样本或总体的问题，只根据合计 100 个数据很可能会得出这样的判断。

但请仔细考虑一下。我们能够断定上面 2 个平均值的差异不只限于样本，而是总体（例如该糖果的整体市场）也“存在差异”吗？还是说如果选取其他样本，就会发现这个差异只是靠不住的“误差”呢？

像这种使用样本数据来确认总体的平均值差异也具有“显著性”的统计方法就是“检验”。我们应该知道，这只是使用样本数据来推断未知的总体数据的特征。

检验的对象除了平均值的差异之外，还有多种应用模式。篇幅所限，这里很难涉及所有内容或者深入介绍详细理论，有兴趣的读者可以参考相关书籍。

检验总体平均值差异的步骤如下。

①提出原假设（归无假设）：“样本平均值与总体平均值之间不存在差异”

②确认能否通过“t 检验”否定（驳倒）归无假设

③如果原假设被否定，说明“二者之间的差异具有显著性”；如不能否定，则“不能认为二者之间的差异具有显著性”

左右检验结果的条件，主要包括样本数据的个数（样本数量）和样本数据的分布方式（标准差）。了解详细的理论当然也很重要，不过我们首先要知道结果取决于这些变量。计算过程用 Excel 就可以完成。

如果问我在实际工作中，用过多少次检验呢？答案是几乎接近于零。一般从事实际工作的人当中，很少有人能够理解“使用样本数据对总体平均值进行检验，其结果是假设被否定，二者之间的差异具有显

著性”的含义。如果分析结果的意义和机制无法获得对方的理解，就很难在工作中实现其目的。

再说，只从理论上指出“未见到统计学上具有显著性的差异，无法用来参考”，并不能解决工作中“那么到底应该怎么做”的问题。

不过了解推断性统计仍然具有以下两个方面的意义。

①能够帮助我们了解自己所使用的数据范围的局限，并在此基础上进行分析（可以防止毫无防备地用现有的所有数据单纯地得出结论）

②与外部分析人员交流时，经常会碰到“显著性”这个词。了解“统计学上具有显著性”的含义，可以准确地理解分析的结果，并与外部的分析人员进行沟通

从这两点来看，了解推断性统计对于提高我们的分析技能、增加相关知识具有重要意义。

后　记

本书介绍了如何将单独的数据分析方法组合起来，形成一个完整的故事，从而解决问题。大家觉得这种方法怎么样呢？

说到数据分析，很多书籍或讲座都会针对各种方法介绍基本理论和应用事例。但在实际工作中运用这些方法时，我们却会发现它们各有利弊。而且很多实际业务问题也都过于复杂，无法用单独的方法去解决。

了解基本的分析方法和思维方式，针对具体问题加以必要的调整，然后才能使它们得到充分应用。为此，需要处理大量实际问题来积累经验。

我在日产一直工作到 2014 年，这期间学到的知识对我现在的企业培训或者对实际业务的支援仍然有很大帮助。这些知识已经成为对任何行业和岗位都通用的软技能。

希望读者在看过本书之后，能够从在日常工作中比较熟悉的、容易解决的问题入手，逐渐培养用故事解决问题的能力。

最后，我想对过去曾经在工作中给予我帮助的各位领导和同事表示由衷的感谢。同时，如果本书能够证明日产公司在解决问题方面的不懈努力和雄厚实力，我将不胜荣幸。

此外，我还要借此机会，对为我提供写作机会的商务杂志《日经信息战略》副主编川又英纪先生表示感谢，也对总是为我增添力量的

儿子优基、女儿朋佳和妻子明子说一声“谢谢”。

DATA&STORY 公司总裁
柏木吉基

出版后记

随着“大数据时代”的到来，数据已经渗透到每一个行业和所有业务职能领域。用数据解决实际问题，也已经成为每一位商业人士必不可少的能力。

然而，对于不具备专业数学及统计学基础的大多数人来说，纷繁杂乱的数据很难发挥出其应有的作用。

面对数据，很多人不知应该从何入手，选择放弃使用这一强大的武器；也有人花费大量时间和精力制作出漂亮的图表，却并不能解决任何实际问题，也经不起推敲和质疑。

本书就是专门针对上述读者而写的。全书没有任何高深的理论和原理，满满的都是从实际工作中得来的经验、技巧和方法。柏木吉基先生长期致力于对数据价值的深耕和挖掘，创造出以解决实际问题为导向的一整套数据分析流程。

本书介绍了在多种工作场景中应用数据的方法和技巧，用最简便的分析方法将数据与图表、公式、逻辑融为一体。从数据采集、处理分析到实际应用，读者应用本书提供的方法，可以轻松地实现数据与实际问题的深度融合。

无论你的工作属于营销、生产还是售后服务，无论分析消费者行为、预测产品销量还是确定工作目标，数据都可以大显神通。希望本书成为你升职加薪的助推器！

服务热线：133-6631-2326　188-1142-1266
读者信箱：reader@hinabook.com

后浪出版公司
2018 年 5 月


人生新算法
用人工智能解读时间、幸运与财富
データの見えざる手
ウエアラブルセンサが明かす
人間・組織・社会の法則
[日]矢野和男——著 范欣欣——译
后浪
江西人民出版社
全国百佳出版社

人生新算法
用人工智能解读时间、幸运与财富

[日]矢野和男 著；范欣欣 译

版权信息

书名：人生新算法：用人工智能解读时间、幸运与财富
作者：[日]矢野和男
译者：范欣欣
书号：ISBN 978-7-210-10466-7
版权：后浪出版咨询（北京）有限责任公司

目录

前　言

第 1 章　时间能否自由使用

1.1　人类行为有规律性吗

1.2　能否根据主观意志自由利用时间

1.3　支配万物的能量守恒定律同样适用于人类

1.4　通过“Life Tapestry”可以俯瞰人生

1.5　从胳膊的活动次数中发现惊人规律

1.6　递降分布统治社会之谜

1.7　反复移动就会出现 U 分布

1.8　我们在各个时间点之间调配“胳膊运动”

1.9　即使不知道微观状况也能预测宏观状况

1.10　时间的利用方法受到规律限制

1.11　“经常动的人”=“有工作能力的人”吗

1.12　把握各频带的活动预算，充分利用所有频带

1.13　没有干劲是因为活动预算用光了吗

1.14　熵是什么？是表示杂乱的量吗？

1.15　自由的牢狱——正因为自由人类才遵守规律

1.16　人类活动的极限可以用热力学公式表示

1.17　人的自由与限制

第 2 章　测量幸福

2.1　能否用科技控制人类的幸福

2.2　幸福的心理学——“积极心理学”

2.3　提升员工的幸福感有利于提高公司收益

2.4　传感器可以测出幸福感

2.5　解读行动中隐藏的符号

2.6　休息时活跃的对话有助于提高生产力

2.7　身体运动会传染，幸福也会

2.8　身体运动活跃的职场的优点

2.9　打造活力职场是一项重要经营项目

2.10　我们也要考虑 IT 会降低生产力

2.11　通过幸福科技创造幸福指标

第 3 章　求“人类行为的方程式”

3.1　人类行为中存在方程式吗

3.2　方程式究竟是什么

3.3　与人的再次见面遵循普遍定律

3.4　以见面概率为基准思考，则时间的流逝各不相同

3.5　1/T定律也适用于回邮件等其他行为

3.6　行动持续越久越停不下来

3.7　1/T定律与 U 分布相同

3.8　记述人类行为的方程式
3.9　将主观概念转化为客观数值
3.10　测量最优体验=心流

第 4 章　认真面对运气

4.1　偶然是不可控的吗
4.2　运气源于与人的相遇
4.3　将运气和相遇转化成理论和模型
4.4　到达度真的是衡量运气好坏的指标吗
4.5　运气好的人在组织中处于什么位置
4.6　领导的指挥能力与现场的自律并不矛盾
4.7　通过数值化，从语言的束缚中解放出来
4.8　通过控制“到达度”，成功实现组织整合，防止开发延迟
4.9　要想抓住运气，对话质量也很重要
4.10　对话即为身体活动的投接球练习
4.11　有关单向交流和双向交流的研究
4.12　根据身体运动的测定值，可以明确定义对话的质量指标
4.13　从“运气也是实力的一种”到“运气即实力”

第 5 章　撼动经济的新“无形之手”

5.1　社会能否科学化
5.2　从科学角度来看，不知“买”为何物
5.3　如何从科学角度解读经济活动
5.4　购买行为的全貌测量系统

5.5　计算机 VS 人类，通过提高销售一决胜负

5.6　学习型机器大显神威的时代

5.7　人类的假说验证分析不能用于大数据

5.8　学习型机器会提高人类“从过去学习的能力”

5.9　3 种人工智能

5.10　通过大数据获取利益的 3 项原则

5.11　学习型机器可用于解决所有社会问题

5.12　人类和工作将与机器共同进化

5.13　人类应做之事与不必做之事

5.14　新的“无形之手”将为世界带来新的“财富”

第 6 章　社会和人生的科学将带来什么

6.1　在濑户内海的直岛描绘未来

6.2　以社会为对象的科学迅速发展

6.3　将服务与科学融为一体的数据呈指数增长

6.4　重大挑战“直岛宣言”

6.5　直岛宣言

6.6　总结——人类旺盛的生命力

后　记

参考文献

出版后记

返回总目录

前　言

2006 年 3 月 16 日是一个分水岭,从那天起，笔者的人生发生了巨大的变化。

当时，我的研究团队正在研究传感器技术及其应用，该技术用于测量并记录人类行为和社会现象。其中一项研究是可穿戴式腕式传感器，它可以持续测量左手的运动。首部样机于 2006 年初开始试用。

该传感器最大的特点是能够 24 小时持续记录人类的行动。然而，没有人愿意作为实验的小白鼠，无死角地记录自己的生活。于是，身为研究组长的我自告奋勇地充当起了小白鼠。

从那天起，一天 24 小时，一年 365 天，过去整整 8 年里，我的左手腕上一直佩戴着记录左手运动的传感器。它 1 秒钟可以测量 20 次，经过日积月累，电脑中存储了详细的加速度数据。利用这些数据可以做出各种解析，比如在过去 8 年里，在睡觉的什么时候翻身了，什么时候集中精力工作了，等等。

短期数据仅仅代表了左手的运动，其意义微乎其微。但是，正如本书所要介绍的，我们花费了 1 周，1 个月，1 年、2 年，经过测量得出很多人的数据，随着数据的不断累积，我们逐渐明白了这项技术的意义何其重大。

继腕式传感器之后，笔者的研究团队走在世界前沿，不断开发用于测量社会现象和人类行为的新传感器技术及其解析技术。在还没

有“大数据”一词的年代，我们就已利用可穿戴式传感器测量社会现象和人类行为，并分析了大量的数据，据此发现了有关人类行为和社会现象的种种秘密，引领了世界的发展。本书对此项研究做出了全面的总结。

回顾历史我们会发现，大到宇宙小到生物，人类针对各种各样的自然现象，构建了以物理学为代表的定量而精密的科学体系，这成为推动 20 世纪社会与产业发展的巨大原动力。

然而，我们再看一下社会现象和人类行为就会发现，虽然社会科学等知识体系在不断发展，但是与物理学等定量而精密的硬科学相比，却依旧停留在定性的层面上。

利用上述传感器技术能够获取大量的数据，通过活用这些数据，我们可以确立针对社会现象和经济活动的定量硬科学体系，进而实现科学预测与调控。

并且，这不单单是科学知识，还直接关系到企业的利益。在本书中，笔者通过呼叫中心、店铺等具体事例告诉读者：基于测量数据来调控人类与社会，将为企业的业绩带来重大影响。

再者，与人类和社会相关的大量测量数据，可能会为我们揭晓人生中一些根本问题的答案，比如“如何提高幸福感”“怎样拥有好运”等。

也许大家会认为它们是哲学和宗教问题，但本书将告诉大家，这类问题也可以从科学角度来解答。

如上所述，本书以从未在科学角度解读过的事物为对象，彻底地实践了科学的方法论。物理学的概念和工具一直用于研究自然法则，但出乎意料的是，它们还能在了解企业利益和人类共鸣方面发挥威力。这是迄今为止的书籍中从未涉及的内容，对其进行阐述正是本书的特色。

各大科学领域并非自古就存在。近百年来，人们扩大了科学的疆界，不断开辟新的领域。近十年间，在《自然》和《科学》等一流科

学杂志上，也开始出现利用定量数据研究人类行为和社会现象的论文。

从这个意义上来说，本书笔者以当事人的身份，真实生动地描绘出了正在推进的科学地平线。

同时，探讨如何以科学为依据来管理组织，也是本书的目的之一。其中还涉及一些管理方面的内容，希望能为每天与业务和组织管理做斗争的管理者和知识工作者带来一些启发。这两者能否共荣共存，全凭读者判断，但是如果本书能激发读者的思考，并为振兴日本经济带来些许启示的话，笔者将不胜荣幸。

2014 年 6 月

矢野和男

第 1 章
时间能否自由使用

1.1　人类行为有规律性吗

在本书开篇请先思考一个问题:人类的行动中存在科学规律吗？我想问的是，你的行动是否遵守了某种科学规律。之所以这么问，是因为这个问题的答案，与能否使用大数据科学调控社会现象及经济息息相关。

迄今为止，从宇宙的起源到物质的结构，人类都是依靠科学来理解世界的。很多时候，进步的契机就是新测量数据的获取。

我们之所以相信宇宙是有起源的——“宇宙大爆炸理论”，相信微观世界的 1 个电子能同时存在于不同的地方——“量子力学”，是得益于数据，这些数据分别源自能在宇宙中检测出微弱电波的天线和能检测出单个电子的测量器。

近年来我们获取了大量有关人类与社会行动的数据，从而使发现与人类相关的新科学和科学规律的可能性大大增加。从有关人类与社会的大量数据中可以推导出规律，并且人们期待着能够利用这些规律，进一步为社会提供正确导向，为经济带来巨大活力。

但是，另一方面，我又有想要否定这一点的冲动。人，难道不可以通过其每时每刻自由的意志与思想，随心所欲地行动吗？难道不可以仅根据自己的意志和喜好来约束行动，而不受规律之类的限制吗？如果真是这样，那么即使我们握有大量的数据，也只不过是对过去的记录，不会直接对未来发挥作用。

人类和社会中，到底有没有普遍的规律呢？面对大量人与社会的相关数据，采取的立场不同（是否认为有普遍规律），对事物的看法就会截然不同。因此，我想先讨论一下这个前提。

1.2 能否根据主观意志自由利用时间

在讨论人类的行动是否存在规律之前，本章将首先聚焦于人类如何利用时间。也就是说，我们想重点探讨一下，人是按照主观意志自由决定如何利用时间的，还是在某些规律的限制下利用时间的。之所以要重点探讨这一点，是因为有时在工作和个人生活中，如何利用时间是头等大事。后面也会讲到，我们利用传感器技术得到了有关人类如何利用时间的大量数据，从中找到了这个问题的答案。

古往今来，很多思想家都论述了有效利用时间的重要性，笔者本人就受到了 19 世纪瑞士的哲学家卡尔·希尔逖的影响。他在其著作《幸福论》中，用整整 1 章介绍了使用时间的方法。而在现代，史蒂芬·柯维在其畅销作《高效能人士的 7 个习惯》中，用 1 章的篇幅阐述了时间管理的问题；管理学的泰斗彼得·德鲁克也在其著作《卓有成效的管理者》中论述，要想成为高效的管理者，最重要的是分析并改善自己利用时间的方法。除此之外，在每年各种各样的文章和书籍中，如何利用时间的问题一直被反反复复地讨论着。

就这样，人们逐渐认识到时间利用问题的重要性，但也仅限于幸福论和自我管理的范畴，并没有将其作为科学研究的对象。各位读者

朋友也是这么认为的吧。

但是，我在这一章的论述恰恰否定了上述内容。今天你在哪方面使用了时间，其实不是随心所欲的，而是在无意识中受到了科学规律的制约。

假设今天你想做 3 件事，想必很多人会在这一天开始之前，列出 To do 清单吧。可能你会认为，在这 3 件事上分别花费多少时间，可以由主观意志决定。

然而，从后面将要介绍的科学规律来看，你并不能随心所欲地使用时间。即使你想要随心所欲，并为此做好了计划，实际也不会按计划进行。如果你回顾一下自己的亲身经历，就会发现有不少这样的情况。当你了解了本章所述的科学规律，就能够理解是什么造成了计划无法照实推进。此外，我还想告诉大家：我们可以依据这一规律，科学地调控自己的时间。

1.3 支配万物的能量守恒定律同样适用于人类

科学分为各种各样的领域，在诸多领域中都存在记述各种现象的基本方程式。

以物理学为例，物体的运动遵守牛顿方程，电磁学现象遵守麦克斯韦方程，量子现象遵守薛定谔方程。想必很多人都听说过这些名词。

但是很少有人知道，这些表示物理现象的方程式说的都是同一件事。实际上，这些方程式都是从守恒定律，也就是从可以保存能量和电荷等的规律中派生出来的。

恐怕没有人能顺畅地写出这些方程式，但是，如果我们知道物体运动、电磁和量子的能量是用什么公式表达的（实际上这些都能用简

单的式子表示出来），就可以当场导出这些方程式。

这些方程式是自然法则的基础，如果它们都是从守恒定律，尤其是能量守恒定律中派生出来的，那么毫无疑问，“能量”的概念才是科学理解自然现象的核心。

事实上，“能量”改头换面，与利用时间的方法产生了联系。一天中可以使用的总能量及其分配方法受到规律的限制，因而你无法按照自己的主观意志自由地使用时间。

在你周围发生的所有现象和变化，都需要能量。能量以各种各样的形态蓄积起来，与所有现象都相关。不仅有原子能、化学能量，还有热能、电能等。

从表面来看，宇宙也好，地球也好，都时刻处于变化之中。虽然能量的形式千变万化，但其总量是一定的，既不会增加也不会减少。

那么，世界为什么会变化呢？我们看见的所有变化，实际上都是从一种能量到另一种能量的转化。例如，苹果从树上掉下来时，苹果的重力转化成了苹果的动能，但是其总能量没有丝毫的变化。也就是说，发生变化的是能量的分配。

相反，只要改变了能量分配，就一定会产生变化。例如，不给低处的物体施加力量，它就不会自动升到高处。因为从低处升到高处需要新能量的产生，而不改变能量分配就无法实现这种变化。从能量分配的观点来看，我们可以用科学明确“能发生”和“不能发生”的事情。

这 300 年的物理学历史，归根到底是一句话：所有自然现象都可以用能量分配这个统一的原理来说明。

如果将人类作为研究对象，情况就复杂了。因为人类有意志、有想法、有情感，这些都会对行动产生影响。明明自然的变化源于能量分配的变化，难道有什么特殊情况能让人类搞特殊化，脱离能量分配原理的限制吗?

1.4 通过“Life Tapestry”可以俯瞰人生

下面介绍一下我们做的实验。

目前我们已经可以使用最新技术，以毫秒为单位测量并记录人类24小时的行动，包括身体运动、与人见面和位置信息等。这10年来，笔者一直致力于开发这种测量技术，并利用该技术来获取数据。

这里的实验使用的是腕式可穿戴传感器（称作“生命显微镜”[1]），可以用加速度传感器测量并记录胳膊的运动（卷首插图1）。高精度的加速度传感器可以每50毫秒测量一次（每秒20次），连胳膊的微小动作也能捕捉到，并将其作为加速度（由于测量的是空间中的3个方向，因此有3个量）记录在存储器中，即使不充电也能连续运转两周左右。我们使用该可穿戴式传感器，分别对12名实验者的胳膊运动进行了4周的测量，一共记录了长达9，000个小时的数据。

所有人类行为都伴随着胳膊的运动，所以该传感器测量并记录胳膊的运动，可以说能够反映人类的生活。简单来说，人睡觉的时候胳膊是静止的，只有翻身的时候胳膊才会动。人醒着的时候，几乎不会静止不动。看一下加速度的记录，什么时候睡觉什么时候醒着便一目了然。

现在为了简单起见，我们只关注胳膊1分钟动几次。从人1天的平均活动次数来看，醒着时，胳膊1分钟平均运动80次；走路时，胳膊1分钟运动240次。相反地，在电脑前浏览网页时，运动次数下降到1分钟50次以下。总之有这样一个特点：不管是什么行为，胳膊在1分钟内都会动几次。

在意识到上述运动特点的基础上，再来看自己胳膊的运动记录，就能很清楚地想起，过去的每时每刻自己都在做什么。如果长期持续记录的话，就可以像看画卷一样，对自己的人生一览无遗。我们用一

幅名为“生命织锦”（Life Tapestry）的图来表示 24 小时的行动：运动活跃时用红色表示，运动较少时用蓝色表示，运动不多不少时用灰色表示。“Tapestry”是织锦的意思，而我们每个人的生活，就像一幅织锦。

我们看一下实际数据（卷首插图 2）。如图所示，我们可以通过 Life Tapestry 俯瞰 4 名实验者一天 24 小时、一年 365 天的生活。能如此这般地俯瞰人生，不失为一种新鲜的体验。实际上，我从 2006 年 3 月起，每天都 24 个小时戴着这个生命显微镜（洗澡和游泳时除外），持续测量胳膊的运动，至今已超过 8 年。现在我把自己的部分人生也一并展示给大家（卷首插图 3）。从图中可以看出，我生活的时间段每年都有几次大的变化。这是因为当时我身在国外，存在时差。

我们可以一目了然地看到，生活模式因人而异。蓝色表示的是几乎没有运动的睡眠时间，有的人每天都很规律，也有人随心所欲，毫无规律可言。每个人的情况都不同，比如性格的差异，工作和家庭情况的差异等，都会让生活模式多种多样。

我们还可得知，时段不同，活动也不相同。在 Life Tapestry 上，红线几乎覆盖了每天早、中、晚 3 个时刻，这表示他们每天都在重复早上上班、中午休息、晚上下班等 3 项活动。此外，也有从动作少的蓝色过渡到绿色的时段。前面也提到过，如果是在电脑前安静地浏览网页，或者在会议上静静地听取他人的发言（或者睡觉），运动就会减少，颜色将由绿色转为蓝色。

综上所述，每个人的活动方式是不相同的，即便是同一个人，每天、每个时段也在进行各种各样的活动。但想必大家会认为，在一天的某个时间采取的某种行动，是根据自己的意志和喜好决定的。

1.5　从胳膊的活动次数中发现惊人规律

真的是这样吗？为了对此进行确认，我们试着用 Life Tapestry 之外的方法来呈现数据。我们以 12 名实验者为对象，统计了 2 周的胳膊运动数据。如图所示，在一定范围内，数据呈直线分布（图 1-1）[2]。

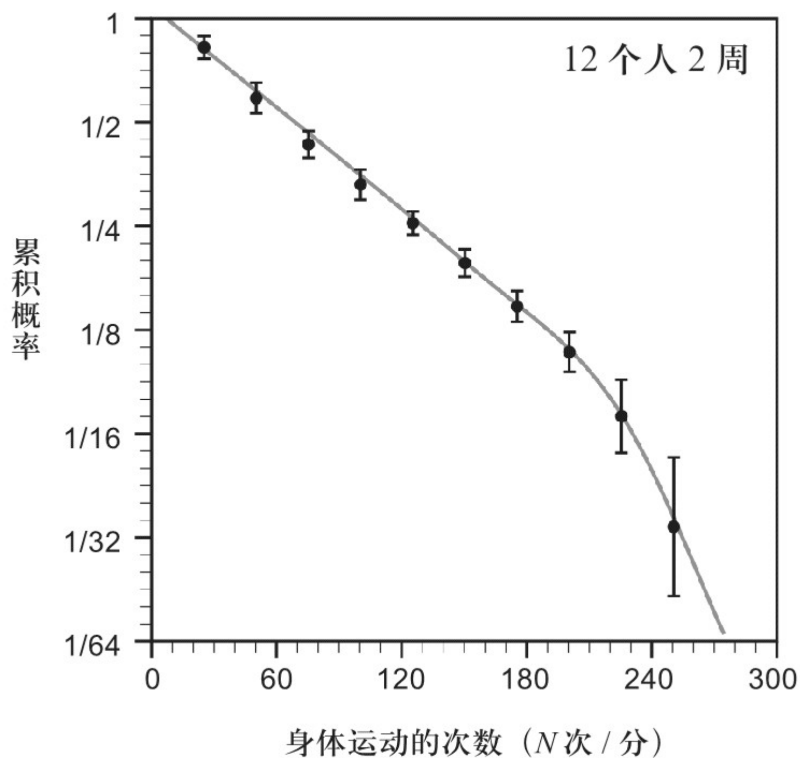


图 1-1 平均每分钟身体运动的次数（N）的分布（12 个人的活动次数×2 周的数据）。纵轴表示累积概率，是由N次以上的身体运动的累计观测时间除以观测总时间得出的。睡觉等身体不活动的时候不进行测量。由于线的倾斜度因人而异，所以结合不同的倾斜度将累积概率进行了规则化。纵轴上表示的 1/2、1/4 等是人们的平均情况。一般用 $1/a$、$1/a^2$等表示，a根据每个人、每一天的不同，在 1.5～3 的范围内变化。平均值为$a = 2$。

该分布图中，横轴表示平均每分钟的活动次数（N次/分），纵轴表示对测量期（2 周）内超过N次的激烈运动进行观测的比率（N次以上的激烈运动的累计观测时间除以观测总时间），这一比率称为“累积概率”。累积概率在纵轴上表示观测到N次以上的运动的概率。例如，横轴数值为 200，纵轴数值为 1/8，就表示每分钟 200 次以上的运动的观测概率是 1/8。

图中，纵轴的刻度都是“对数”。刻度数值等距分布，为倍数关系。这样绘图是为了引出后面将介绍的“U 分布”——一种人类行为和社会活动中常见的统计分布。这里的 U 是指 Universal，即“普遍性”的英文首字母。当统计呈现 U 分布时，如果我们采用“半对数作图”进行绘制，图形就会呈直线，因此一目了然。这与数学上“指数分布”的特点相类似。

U 分布是我们在广泛测量人类行为和社会行动时发现的。如您所见，人类胳膊的运动也呈 U 分布。简单来说，特点如下：如果测量期长（超过 1 天），那么低于 50 次/分的平稳运动多，激烈运动少。其降低趋势遵循指数函数（章末注 1）。

这一倾向非常规律且典型。每分钟超过 60 次的运动占 1 天的一半（1/2）左右，但是，每分钟超过 120 次的运动会减少一半（1/4）。而每分钟超过 180 次的运动，会进一步减少一半（1/8）。如果把这种情况反映到图表中，那么在半对数线图中将呈现一条逐渐下降的直线。

在科学领域，物质中原子和分子的热能分布状况也与 U 分布相同。不管是空气还是水，世界上所有物质的原子都在热能的驱动下不停地运动。例如，空气中约有 1/5 的氧分子，这些氧分子在热能的作用下，分别朝不同的方向运动。这些运动不仅方向不同，速度也各不相同——既有高速运动的氧分子，也有低速运动的氧分子。如果横轴为氧分子（一般来说是分子和原子）的动能，即热能，纵轴为超过该热能数值的氧分子的频率（累积频率），结果仍会呈现相同形状的分布，从半对数线图来看还是一条直线。以热能为横轴的统计分布称为“玻尔兹曼分布”，这是一种决定了所有物质的热能性质的基本分布。

我们最近发现，若继续使用传感器测量并研究人类的身体运动，就会明白即便横轴不是原子的热能，而是各种有关社会现象的量，结果还是会呈现相同形状的分布。比如胳膊每分钟的活动次数、顾客在店铺的货架前停留的时间等。

令人惊讶的是，有关人类行为和社会现象的大量统计分布图都呈现这一形状。更让人感到意外的是，这种现象如此频繁多见，却没有人察觉。不仅是原子能领域，在人类行为和社会现象中也普遍存在这一分布规律，因此我们构建了这一规律的理论依据，并将该数理统计的分布命名为“U 统计”。

有关胳膊运动的实验结果让人感到不可思议。我们测量并统计了一天内胳膊每分钟的活动次数后发现，结果与图 1-1 一样，是呈直线形的 U 分布。我们认为这也许是偶然现象，于是又在另一天进行了实验，结果还是呈现 U 分布。此外，我们还根据 12 名实验者的数据绘制了图表，结果发现他们每天的数据全部呈现 U 分布。鉴于结果太过神奇，我也调查了一下我自己的数据，可每天的结果仍然呈现完美的 U 分布。

再琢磨一下这其中的意义，我们会发现更加惊人的事实。被实验者都认为自己可以根据自己的意志和想法来决定行动，因此每个人发出的行动不同，其动作也各有特点。而这正与具有普遍性的 U 分布相矛盾。

举例来说，人在发言时的运动频率是 150 次/分，浏览网页时的运动频率是 30 次/分。如果人们按照自己的意志选择这一天何时采取何种行动，由于选择不同，呈现的分布规律应该也各不相同，因此不可能每天都呈现统一的 U 分布。更何况，每个人的工作不同，性格和年龄也不同。可神奇的是，人们都像被施了魔法一般，一天 24 小时都沿着 U 分布行动。

这简直像变魔术一样，不管怎么洗牌，最上面的一张都是黑桃 A。但是，现实的人生和世界不是魔术，没有什么机关窍门可言。到底是什么在无形之中引导着人类的行动呢？让我们再深入研究一下。

1.6　递降分布统治社会之谜

这种普遍性递降的统计分布是深入理解人类行为、社会及经济现象的关键。下面我将具体介绍一下这个重要的统计分布曲线。

在统计学中，以正态分布——“吊钟形”分布居多（图 1-2）。其特点是，以平均值为中心，两侧（高于平均值和低于平均值）的曲线左右对称。

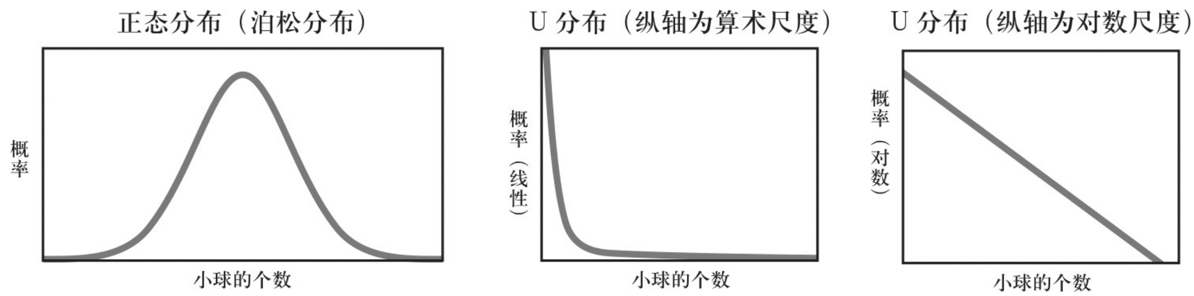


图 1-2 正态分布（泊松分布）和 U 分布的比较。正态分布呈吊钟形，以平均值为中心，两侧（高于平均值和低于平均值）的曲线左右对称。U 分布是逐渐下降的曲线分布。半对数线图则呈一条直线分布。

我们来举一个正态分布的实例。掷了几次骰子后，点数的平均值是多少？假设我们掷了 5 次骰子，点数依次为 3→5→1→6→1，那么 3+5+1+6+1=16，16÷5=3.2，平均值就是 3.2。如果我们反复掷多组，每组还是掷 5 次，虽然得出的平均值各不相同，但结果均以 3.5 为中心（骰子点数的平均值是 3.5）上下波动。这就是正态分布。掷 5 次骰子得到的平均值大多都接近 3.5，很少会出现像 1 或 6 这样偏离 3.5 的数值。因为想让平均值为 1，就必须每次都掷出点数 1，而这是极其偶然的情况。

在统计学的相关书籍中，很多书的内容都是以正态分布为前提的。想必许多人都认为，世界上大部分现象都可以用正态分布表示，其余的情况都是例外——“正态”这个名称本身就包含了这层意思。

但是，在现实社会中，大数据的统计分布大多为递降的 U 分布。在这一分布中，当变量为 0 时频率最高，随着变量的增大，频率会逐

渐降低。并且，正态分布与 U 分布的形状差别很大，前者呈吊钟形，后者呈递降曲线（章末注 2）。

我们该如何理解这一现象？难道只是碰巧变成了一条递降曲线？我查阅了许多书籍和文献，都没有找到人类行为呈递降分布这一规律的原因。每次见到统计学和物理学界的学者，我都会问：“为什么统计学中用正态分布进行研究，而物理现象用递降分布研究呢？”明明是基本的问题，却几乎没有人能回答上来（唯一明确回答我的，是巴西统计物理学界的权威康斯坦丁诺·查理斯教授。他的回答令我深受启发，在此也向他表示感谢）。很长一段时间以来，我的眼前仿佛烟雾弥漫，一片模糊。但是最近我通过模拟和解析，终于找出了让人心悦诚服的答案，眼前也顿时云开雾散，一片清明。

“反复之力”，是一种我们平时没有感知到的力量。通过研究，我渐渐看到了这种力量是如何撼动整个社会的。

1.7　反复移动就会出现 U 分布

大数据中常见的递降 U 分布的本质是什么？下面将用图像直观地进行介绍[3]。

首先，我们看一下由 30×30（900 个）的方格构成的网格图（图 1-3）。假设有 72,000 个小球，我们将其完全随机地放到图中[4]。

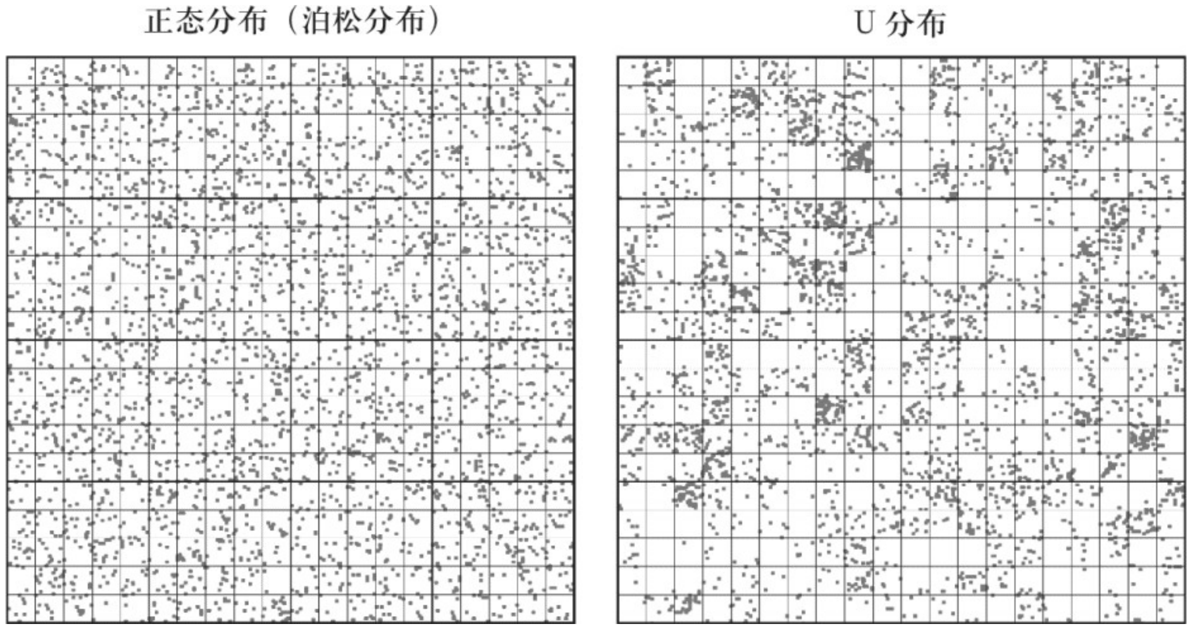


图 1-3 网格中散布的小球在正态分布和 U 分布中的差异。本书是用 30×30 的网格实验进行解说的，但为了清晰起见，我们只扩大展示了其中一部分。并且，要展示所有小球的话会有重叠，所以我们把每 10 个小球整合为 1 个小球显示在图中。

如果用电脑模拟实验，那么可以随机生成小球的位置。首先生成水平方向（x）和垂直方向（y）上 1～30 个随机数，然后把小球放在（x，y）的位置上。这样一来，一个方格中平均会有 80 个小球，80（个）×30（格）×30（格）= 72,000（个）。

这张网格代表了你的 1 天，每个方格代表 1 天中的 1 分钟。网格图中的方格总数是 900，假定 1 天有 900 分钟（15 小时）的活动时间，则方格总数与活动时间相对应。此外，方格中的小球个数代表胳膊 1 分钟的活动次数。每个方格中平均有 80 个小球，即假定胳膊 1 分钟平均活动 80 次。现实生活中，虽然活动时间和胳膊的平均活动次数会因人和状况的不同而有所变化，但是也会出现 1 分钟 80 次的活动次数（即使换一个数字，也不会对以下结果造成影响）。

如图 1-3 所示，虽然每个方格中的小球数量不同，但是基本平均在 80 个左右。这一统计分布即为正态分布（专家称之为“泊松分布”，以区别于正态分布，但是两者基本相同，因此本书不作区分）。

通过掷骰子的方法也可以得到基本相同的结果，只不过花些时间罢了。在每个方格中掷 23 次骰子，得出的点数总和基本以 80（准确说是 3.5×23=80.5）为中心上下波动。这和刚才的模拟结果相同，呈现正态分布。

现在我们只是随机分配小球，每个方格中的小球不会自主地从一个方格移到另一个方格。接下来，我们在方格之间移动一下小球，看看情况如何。

我们随机选择两个方格，将其中一个方格中的 1 个小球移到另一个方格中，然后进行反复移动。大家可能觉得，原本就是随机放置的小球，方格也是随机选择的，即使移动小球，结果也不会有所变化。我曾经给很多人出过这道题，所有人的回答都是“结果不会变”。

但是，事实胜于雄辩。请看一下图 1-3 中位于下面的图。这张图是“反复移动”10 万次后的结果。反复移动的次数越多，小球的分布就越趋向“斑点状”。其实，因移动而产生的“斑点状”才是现实社会的大数据中常见的递降 U 分布。也就是说，我们将小球按照从多到少的顺序，统计其数量分布情况，会发现数据呈现 U 分布。U 分布的制作方法是非常简单的。

与 U 分布相比，原来的正态分布整齐划一。从结果来看，两者的差异很明显，正态分布是随机的、均匀的，而递降分布是散乱的“斑点状”——比彻底的随机还要散乱。这话听起来有点矛盾，但事实就是如此。其实，基于同一随机数的随机性，产生的是十分均匀、整齐的状态。而 U 分布中允许出现“不均匀”，是一种更自由的状态。

U 分布是将玻尔兹曼分布普遍化的产物，但即使是专门研究物理的人，恐怕也是第一次如此直观地看到玻尔兹曼的空间分布。至少笔者在任何一本统计物理学的书籍上，都没有见过这样的图。翻开统计物理学的书籍就会发现，玻尔兹曼分布公式（以温度的倒数为指数的指数函数公式）随处可见。但是，只有公式的话，我们完全无法想象其空间分布究竟是怎样的。

气体中的分子之间经常相互碰撞，与此同时会交换彼此的能量。这类似于方格之间小球的反复移动，因此我们也就不难理解，分子能量的分布也与 U 分布相同，是呈递降的玻尔兹曼分布。

从结果来看，递降 U 分布中，小球集中于少数几个方格中。定量分析的话，在前 30% 的方格中，小球的数量占总体的 70%。我们经常谈到“二八定律”，即前 20% 中集中了整体的 80%。例如，人们经常说 20% 的员工取得了 80% 的销售，20% 的企业创造了 80% 的 GDP 等。虽然这个 U 分布没有完全集中到 20/80 这种程度，但也已经相当集中了。

那么，小球的分布呈现斑点状意味着什么？打个比方来说，是方格和方格之间产生了“贫富差距”。分配小球时，自然而然就会产生两种方格：一种是集中了很多小球的富裕方格，另一种是没怎么分到小球的贫穷方格。之所以会产生这种差距，是因为小球在方格和方格之间反复移动。

有趣的是，明明每个方格都是“机会均等”的，小球却集中到了少数特定的方格中。也就是说，即使“机会均等”，产生的结果也不平等。即使是均等地“反复移动”，也会产生不平等的结果。

我们必须记住的是，小球集中到特定的方格中，不是由方格自身的特殊性，比如能力差距等导致的，而仅仅因为均等的反复移动。即使我们不做方格之间存在能力差距这种假设，由于概率问题，还是会导致差距。也就是说，“反复之力”造成了这种“贫富差距”。

说句题外话，从经济贸易出现起，自给自足的人类之间就产生了贫富差距这种原始的模型。

我们往往以为凡事必有因。总是以为富裕的人和不富裕的人之间，在行动方面应该存在差异，然后去探求这一结果背后的原因。但实际上，当发生了很多次反复移动，即使没有确切的原因，其结果分配还是会明显偏向一方。我们必须记住，资源（小球）分配不均绝不是因为能力和努力的差异，而是由“反复移动”产生的统计力量导致

的。在现实社会中，不仅有自然产生的分配差距，还存在能力差距，因此贫富差距进一步扩大。

在“反复之力”的作用下，资源分配的差距阐释了人类广范围的行动和社会现象。而将此上升为理论依据的就是 U 分布。

在此，我们需要考虑一个简单的问题——小球的分配。随机反复移动小球的话，会出现怎样的结果呢？这个问题我问过几十个人，其中很多人是理科博士。然而，让人惊讶的是，这么简单的问题却让他们调动了全部的经验和知识，最终也没能预测到结果。很多人回答，小球的分配还是随机的，没有变化。这个结果明确表明，对于包含“反复运动”的现象，我们的预测能力是何其欠缺。人类有一种强烈的倾向，即总想借助因果关系来认识世界。但是，因果关系这种思考方式，可能并不适合预测多次反复后的结果。

1.8　我们在各个时间点之间调配“胳膊运动”

现在，让我们回到之前关于人类 1 天的运动的问题。人类的运动遵循的是正态分布，还是递降 U 分布，又或者是其他的分布？实验结果表明，遵循的是递降 U 分布。那么这意味着什么呢？

首先我们来确认一下，一个人的行动中是否存在很多反复运动和统计要素。

当然是存在反复的。如果将你 1 天的行动按分钟累积起来，那么 1 天大约有超过 900 分钟的活动时间。这是一个很庞大的统计数据。

现在我们再看看图 1-3 中 30×30 的方格。用一张网格图代表 1 天，网格中的一个方格对应 1 分钟。1 分钟做什么可能取决于各种情况，并由此决定了这 1 分钟里胳膊的活动次数。如果方格中有 10 个小球，就表示每分钟胳膊运动了 10 次。

我们每天活动的时间大约有 900 分钟，胳膊活动约 7 万次。将这 7 万次的运动分配到每分钟，如果我们在各个时间点的行动类型都是随机的，那么小球的分配应该会呈正态分布。中心值就是 70,000/900 的平均值。

但实际上，小球的分配呈现“斑点状”的递降 U 分布。这一分布规律的本质，是小球在方格之间的反复移动。类推到人类行为的话，一张网格图相当于 1 天，由 900 个 1 分钟组成，在每个 1 分钟（方格）之间，胳膊反复调配着 7 万次的运动。

“调配”是指在哪个时间点活动胳膊。胳膊 1 天大约活动 7 万次，这一总数基本是固定的，我们只是配合优先活动程度调整了胳膊的运动。

例如，上午减少活动量（胳膊的活动），下午全身心投入工作，向顾客提出方案（频繁活动胳膊）就属于这种情况。又比如，有时候我们会在 11 点之前集中精力按时完成资料制作（频繁活动胳膊），之后稍事休息（减少胳膊活动）。胳膊的活动次数是有限的，不需要优先活动胳膊时就保存体力，而需要优先活动时就多分配一些，这就是“胳膊运动的调配”。或许我们都在下意识地调整行动，这种调整细致入微、不计其数。我们每分、每时、每天都在使行动最优化。

你根据是否需要优先活动胳膊而无数次地调整胳膊的运动，递降 U 分布则正好证明了这一点。如果你停止这种优化行为，那么胳膊运动的分布应该会接近图 1-3 的正态分布，而实际上这种情况不会发生。人类每天都会将有限的胳膊运动资源分配到每时每刻的行动之中。

现在让我们回到本章开篇提出的问题，即人的行动会像物质一样受到能量的限制吗？宇宙中的所有变化都源于能量的转换，唯独人类的行动就取决于意志、喜好和情绪吗？只有人类是特别的吗?

结论就是，人类的行动并不特别。对于人类的行动，我们不能像原子运动和电磁波一样，来严格定义“能量”。但是，胳膊活动次数的

分布与原子能量的分布可以用同一个公式表示。

这并非偶然。因为两者都在反复调配有限的资源，并呈现出相同的结果。

在物质中，热能在分子间被反复调配。而你每时每刻也都在调配胳膊的运动。一个是“能量”，一个是“有限的资源”，虽然名称不同，但本质相同。更准确地说，能量是有限资源的一种，可以视为特殊的资源。

实际上，胳膊的运动是我们活着的重要资源。你身体的运动，也可以说是你得以采取行动的唯一来源。在模拟实验中，我们将该资源形象地比作小球。小球随机地反复移动是再简单不过的模型，我们可以通过这个模型解释人类行为这一复杂的体系。它将表面的宏观现象与其背后微观下的反复状态联系到了一起。下面我们来进一步思考这种宏观和微观的关系。

1.9　即使不知道微观状况也能预测宏观状况

19 世纪前半叶是定量化理解“气体受热膨胀”（物质的基本特性）的时代，这个时期确立了热力学理论。但是，气体受热膨胀现象的背后到底发生了怎样的微观变化？也就是说，分子运动与气体状态的关系是怎样的？实际上当时人们对此并不清楚。19 世纪后半叶到 20 世纪前半叶，从原子运动的角度说明宏观物质性质的理论才确立下来。该理论即“统计力学”。

统计力学通过“反复之力”——构成气体的无数微观分子的反复碰撞，来说明上述气体膨胀的宏观现象，并且将这种反复上升为理论，从而能够预测现象。其基本的理论体系是由詹姆斯·克拉克·麦克斯韦、路德维希·玻尔兹曼、吉布斯、爱因斯坦构建起来的。小到身边的物质，大到宇宙万象，这一理论体系解释了很多现象。

我们从中发现了一个重要原理，即随着反复移动次数的增多，即使不知道具体的微观状况，也能预测并控制宏观现象。我将这种原理称作“多次移动原理”。也就是说，当移动数量足够大时，并不需要了解每次移动各自的具体情况，重要的是与移动相关的少数规则。

没有这个原理，我们将无法预测由无数个原子构成的自然现象。不管出现多么先进的超级计算机，都不可能对与自然现象相关的原子运动的初期条件了如指掌，也不可能模拟所有原子的运动。但是，当存在大量微观的“反复移动”时，大部分微观状况都不会对宏观现象产生影响，而只有极少数一部分信息才会影响宏观现象。借助该原理，自然现象中的微观和宏观就联系到了一起。

同理，对于人类每天超过 7 万次的胳膊运动，即便我们不考虑诸如意识、想法、情绪、事情等时刻变化的具体情况，也能实现科学的预测和控制。如果我们测量人类行为并调查其分布规律，无论被测人持有什么意识、想法、情绪或者做了什么事情，结果都会呈现 U 分布。这也是“多次移动原理”的一种体现。同样地，虽然我们无法控制空气中的单个分子，但是我们能预测并控制由无数个分子的反复碰撞产生的气压和温度。

所谓 U 分布统计，就是将宏观和微观结合起来的理论。在每时每刻的微观行动中，重要的是明确小球的存在，即资源是什么，以及这些小球按照什么规则移动。如果没有移动，结果就会呈现以往统计学中常见的正态分布，而如果发生移动，就会呈现 U 分布。在原子的世界中，能量作为“小球”这一资源发挥着作用，而在人生和社会中，“小球”资源也发挥着同样的作用。从微观上来看，能量的反复转换创造了自然现象。同样地，胳膊每一秒的反复运动也创造了社会现象。并且，我们能够科学地预测这些反复转换或运动。

1.10　时间的利用方法受到规律限制

在如何利用时间方面，我们能从前面得出的结论中获得一些启示吗？答案是 YES。

将能量普遍化的“资源”及其“反复转换”阐释了人类的行动。现实中的所有现象都可以用资源分配的变化来说明。例如，由于能量守恒定律的限制，我们可以断定水从高处流向低处，而绝不会从低处流向高处。实际上，如果人类的行动接受资源转换定律（相当于能量守恒定律的普遍化）的支配，那么人类行为就会受到严格的制约，如此一来，时间的利用方法也会受到严格的限制。

比如说，我正在用电脑写稿，而截止日期已迫在眉睫，那么在这段时间我能一直写下去吗?

写稿时，胳膊的运动是有特点的。如果按照同样的速度写稿，那么胳膊每分钟的活动次数会在某个范围内波动，我们称之为运动的“频带”。假设写稿期间每分钟平均活动 60 次，动作较多时达到 70 次，动作较少时为 50 次，那么运动的频带就是 50 ~ 70 次/分。如果保持每分钟平均 60 次的运动，那么 1 天的运动将呈现什么分布规律呢？答案是以平均值 60 次/分为中心，呈吊钟形的正态分布——竟然不是 U 分布，这怎么可能呢？我们可以断言，在现实中这是不可能发生的。

假设在众人面前发言时，胳膊 1 分钟的活动次数为 120 ~ 180 次（平均 150 次），那么我能够发言 5 个小时，而只写 3 个小时的稿子吗？这显然是不可能的。因为这样的话，运动的统计分布就不是 U 分布了。U 分布是沿着一个方向逐渐下降的曲线，如果活跃的运动比安静的运动花的时间还要长，就违背了 U 分布的规律。在 U 分布中，快速的运动与缓慢的运动相比，花费的时间只能更短。想要使运动呈现 U 分布，就必须减少 120 次/分的运动，或者增加 60 次/分的运动。

如果我只是发言和写稿子，就会产生未被使用的频带。即使没有将低于 50 次/分、70 ~ 120 次/分、高于 180 次/分的运动列入假设，这些运动还是会按照 U 分布分配到一定的时间。稿子急需赶工，虽然我想优先写稿，但还是不得不花相应的时间在那些未被使用的频带上。如果我非要按自己的想法做呢？比如我勉强自己遵守了时间表，但有

可能我的发言一看就没有用心，或者我心慌意乱地走来走去，根本无法集中精力写稿，也就是说时间还是花费在了不同频带的行动上。如此看来，如何有效利用因写稿和发言而未被使用的频带，是关系到时间使用方法的重要课题。

我们以为可以自主决定 1 天的必做事项并为各个事项分配时间，但那根本就是天方夜谭。

1.11 “经常动的人”=“有工作能力的人”吗

U 分布的有趣之处在于，身体 1 天的运动次数的分布，基本由运动总次数（或者是 1 天中平均每小时的活动次数）这个唯一变量决定。如果运动总次数确定了，那么按照 U 分布，各频带运动所花的时间也就确定了。我们称之为“活动预算”。

如果 1 天的活动总量（身体运动的总次数）确定了，那么分配到各频带的活动预算也就确定了，并且不能超出预算时间。反之，无论多忙，各个频带的活动都必须花掉相应的预算时间。再具体点说，从实验中我们得知，低于 60 次/分的活动，必须花费 1 天活动时间的 1/2 左右，60～120 次/分的活动必须花费 1 天活动时间的 1/4 左右，120～180 次/分的活动必须花费 1 天活动时间的 1/8 左右，180～240 次/分的活动必须花费 1 天活动时间的 1/16 左右。

每分钟的平均活动次数因人而异。这种差异也会体现在分布图中。活动次数少的人，递降分布曲线的倾斜度大，呈快速衰减状态；活动次数多的人，递降分布曲线的倾斜度小，呈逐渐衰减状态（在图 1-1 中，倾斜度的差异消失，纵轴呈现正态化分布）。该倾斜度的倒数称为“活动温度”。统计力学中，根据玻尔兹曼递降分布中倾斜度的倒数定义了物体的“温度”，这里是据此进行的类推（章末注 3）。但在此我们不对统计力学的具体理论做深入探讨，只要大家理解 1 分钟

的平均活动多（或者说 1 天的运动总次数多）称作“活动温度高”，反之称作“活动温度低”即可。

当我们注意到这一点后再去看实验结果，就会发现物体分为冷、热两个状态。同样地，人的活动也分为活跃的“热天”和安静的“冷天”。

我们进一步发现，活动温度因人而异。有活动温度偏高的“热人”，也有活动温度偏低的“冷人”。“热人”的平均活动次数约为 120 次/分，而“冷人”的平均活动次数约为 60 次/分。

活动温度偏高的“热人”，平均来说活动得较多。而活动温度偏低的“冷人”则活动较少。乍一看好像活动温度偏高的人更加活跃，能做更多的工作。但是，事情并没有这么简单。

假设活动温度高的人需要做低频带的活动（动作少的活动），比如写稿，即使他们不愿在高频带的活动（动作频繁的活动）上花时间，也不得不这么做。因此，他们无法把时间花在写稿这种低频带的工作上。也就是说，这种人很难长时间伏案工作。

相反地，活动温度低的人（即递降分布图中倾斜度高的人）即使想从事高频带的工作（活动较频繁的工作），也很容易出现活动预算不足的情况。因此，他们很少在这种工作上花时间。

1.12　把握各频带的活动预算，充分利用所有频带

既然人类的运动遵循 U 分布的规律，那么为了有效利用全天的时间，重要的是把握各频带的活动预算，然后均衡地使用所有频带的活动预算（能量）。如果无视这一点，就算我们制定了 1 天的计划，最终还是无法按计划进行。

轻易制定的计划可能有害而无利。之所以这么说，是因为如果我们不知道这个原则，就可能陷入自我厌恶的状态，认为没有完成计划是由自己的意志薄弱所致。

然而事实并非如此。计划之所以没有完成，只不过是因为你在推进必做事项时，用光了各频带所有的活动预算而已。

在这里，笔者借用了无线通信领域的“频带”一词。在利用无线电波的通信领域，一个常见的思考和实施方式,是将活动分配到每个频带（也称作波段），并充分利用所有频带。

电磁波广泛应用于收音机、电视机、手机、GPS 定位系统以及收费处的 ETC 等。一方面，各频带的各个用途之间必须互不干涉；另一方面，如果不用完所有频带就会浪费。因此，按照法律规定，使用电磁波时必须根据用途确定电磁波的频带，并用完所有频带。

然而，对于人类活动的频带，由于之前不知道上述原理，因此没能有效地利用。

如果使用可穿戴式传感器的话，我们不仅能掌握自己的哪个活动利用了哪个频带，还能明确 1 天的活动预算。此外，通过该传感器，我们还能得知一天中某项活动的预算还剩多少。这就像一边开车一边在汽车仪表盘上确认汽油剩余量一样，与不看计量仪比起来，应该能更顺利地到达目的地吧。

1.13　没有干劲是因为活动预算用光了吗

从这一层面来看，现实生活中的我们就像在完全不知道汽油剩余量的情况下开车。这样做会发生什么呢？当然是汽车因为突然耗尽燃油而抛锚。假如是人，可能 1 天还没过完，突然某项活动的时间预算就用光了。

当我们用光了活动预算，会发生什么事情？恐怕我们将无法继续那项活动，或者失去干劲。

总感觉没有继续做下去的干劲了——想必每个人都有过这样的体验。事实上，没有干劲可能是因为活动预算用光了。

在预算用光的情况下，如果我们还想勉强自己继续做下去，结果可能就是睡过去，或者无法集中精力（精力不集中时的动作必定不同于精力集中之时，也就是说用到的是不同的频带，因此可以在不违背 U 分布的情况下持续下去）。说不定这也是造成压力的重要原因之一，目前我们正在推进相关研究。

再者，人有没有动力和干劲，是否会继续做下去，一个很重要的因素是有没有按照 U 分布来分配时间。是脱离 U 分布勉强行事，还是按照 U 分布开展行动，这将产生两种截然不同的走向。

如果能够根据传感器测量自己的行动，将每天已使用完的频带和剩余频带的信息可视化，那么我们就可以在每天的行动中，发现无数个优化时间的机会。

1.14　熵是什么？是表示杂乱的量吗？

人的活动遵守 U 分布，受到身体活动这一有限资源（能量普遍化的产物）的限制。这意味着，我们是在自然法则无形之手的支配下开展活动的。

以能量带来的制约为研究对象的科学体系称为热力学。热力学在人类活动方面也可以成立，这不足为奇。

热力学中存在 3 个基本定律。

热力学第零定律，指系统中存在"温度"的概念。也就是说，东西有冷、热之分。前面说过，人类活动中也能导入温度的概念，也有

冷、热之分。

热力学第一定律，是能量守恒定律。前面说过，宇宙中所有变化的资源——能量的总量是保持不变的。人的活动也受到能量普遍化的产物，即“身体活动”这个总资源的限制。

热力学第二定律，是熵增定律。该定律也适用于人的活动（热力学第三定律也为人所知，但这是很久之后才追加的一条，定律的普遍性也比其他定律低，因此这里不做解释），我们后面再做说明。

先解释一下“熵”这个词吧。

熵这个概念，最初由德国的物理学家鲁道夫·克劳修斯在19世纪后半叶提出，用以表述“热”这种奇妙的物质性质。后来，奥地利的物理学家路德维希·玻尔兹曼明确揭示了熵与原子运动的统计关系。

我们一般认为，熵表示的是对象体系的“杂乱”“无序”和“随机”（实际上这是不正确的，关于这一问题将在后面进行说明）。

我们还进一步明确，这种“杂乱”会一直增加，绝不会减少。打个比方，如果自己的桌子上乱七八糟，无法整理，有人就会找借口说，“熵会增加，无计可施”。

全宇宙的熵也会不断增加，于是有人就悲观地预测，宇宙早晚会越来越乱，最终变成一个完全杂乱、随机的“死亡世界”。

德国生理学家、物理学家赫尔曼·冯·亥姆霍兹最先提出了这个预测，并称之为“热死亡”。但是，这又和现实相矛盾——宇宙从起源至今的140亿年间，形成了地球、丰富多样的自然秩序和生态系统。因此，自古以来，关于熵的解释就争论不断。

实际上，将熵增的世界视作“随机无序的世界”“死亡世界”的观点，是对熵增的很大误解。熵是衡量“杂乱无序”的尺度，因此在很多人的印象中，熵增的世界中都是随机的噪音。在这个由随机数创造的杂乱的世界上，美丽地球的自然秩序和生物活力被毁灭殆尽，只剩“沉默和噪音”。打个比方，以前的电视把模拟信号转换为图像，播放结束

后，屏幕上会出现雪花，也就是变成了只有白噪音的世界。所谓白噪音，指的是所有频率（频带）的噪音均匀重叠在一起所产生的噪音，可以说是一种非常不规则的噪音。这就是最杂乱无序的终极世界——“死亡世界”的样子。

但是，熵一旦增加就会变成白噪音世界的观点是错误的。

只要做一个统计分布图，真相就会一目了然。如果将白噪音世界中杂乱的物质分子做成频率分布图，结果就会呈现正态分布（参照图1-3），而不是U分布。在噪音均匀的状态下，熵比较低。

熵大的状态，指的是更加自由、大胆地分配资源（能量）的世界。而随机均匀的世界，指的是完全相反的低熵状态。也就是说，比起吊钟形的正态分布，U分布状态下的熵更大。实际上，熵最大的分布就是玻尔兹曼分布（U分布），这是要在统计力学课堂上最先学习的基础知识。

在本章的前半部分，我们对比了正态分布和U分布。实际上，图1-3中所示的U分布世界，才是熵很大的状态。比起正态分布，那是一个更加参差不齐的斑点状世界，是从限制中解放出来，自由反复地调配资源的世界。唯一受到限制的是总资源，即总能量。

因此，熵不应该像以往一样被视作衡量“杂乱”“无序”的尺度，而应该理解为衡量“自由”的尺度。至少这样理解不会产生误会。

随着时间的流逝，宇宙逐渐从大爆炸，即诞生时期的藩篱（均匀性）中解放出来，偏离正态，趋向自由。实际上，在思考人类活动的熵时，将熵视作衡量“自由”的尺度至关重要。

1.15 自由的牢狱——正因为自由人类才遵守规律

从数学角度定义熵的，是奥地利物理学家路德维希·玻尔兹曼。

在肉眼可见的宏观世界背后，存在一个看不见的微观世界（原子世界）。在这个微观世界中，有无数个原子在以惊人的速度反复地运动和碰撞。乍一看物质在宏观上没什么变化，微观上却在反复变化着。但是，不管怎么组合，不断变化的无数微观状态（分子、原子的位置和移动速度）与宏观状态下的物质都是一样的。

玻尔兹曼在不改变温度和压力等宏观性质的情况下，计算微观状态下可能出现的组合数量，将其总数的对数作为熵的数学定义[①]。这就是有名的玻尔兹曼公式，它还被刻在了维也纳玻尔兹曼的墓碑上。

这意味着，在不改变宏观状态的情况下，可以实现的微观状态有很多。这与熵大的状态相对应，并且与将熵视作“自由”尺度的理解完全一致。

对于人类活动的熵，也可以直接用这个方法来定义。以人 1 分钟的活动次数表示人的微观状态，在一段期间内，计算微观状态的组合总数，然后取对数，就可以定义人类的熵了。这表示了人类活动的自由度和不受束缚的程度。

假设我们身处全天被强制重复同一机械工作的状况。此时人的行动可选的状态组合数较少，熵变低，即变成不自由的状态。

如果我们每隔 1 分钟用掷骰子来改变行动的话会怎样呢？由于受到每隔 1 分钟就要改变行动的限制，结果接近均匀的随机状态，也就是非自由状态，熵变低。而所谓熵大的状态，指的是从所有的限制中解放出来的自由状态。这实际上就是 U 分布表示的状态。

我们这样定义人类活动的熵，并逐渐延长时间的话，熵也会逐渐增加。

熵会自然而然地增加，意味着在自然规律的推动下，我们和宇宙从各种各样的限制中解放出来，走向自由。

然而，讽刺的是，自由伴随着代价。一直处于自由状态，反而会受到限制。

德国作家米切尔·恩德凭《永远讲不完的故事》和《毛毛》名满天下，他在作品中提到了一个“自由监狱”，讲述了可以自由选择的权利逐渐成为折磨自己的限制的故事。

虽然状况各异，但是一般来说，对自由的认可意味着控制难度的提高。而难以控制的后果，就是人的“活动效率”受限。我们已经分析过了，因为自由，人无法把时间 100% 花在一项活动上，从而也会限制业务生产力和时间利用方法等方方面面。

1.16　人类活动的极限可以用热力学公式表示

在物质世界中，因熵增加出现的效率受限理论已经确立了。从该理论中我们可以得知，熵的增加使发电站和发动机的效率受到限制。

将热能转换成机械能的装置叫热机。核电站、火力发电站和汽车的汽油发动机都是热机。

热机的效率（热效率）有上限这一点已广为人知。例如，在核电站，核裂变产生热能，在热能作用下产生水蒸气，进而带动涡轮机运转，产生能量。但是，由于熵增加的限制，不管技术怎么发展，将热能转换成能量的效率都不会达到 100%。该转换效率与能量的成本直接相关，因此存在极大的经济价值。

虽然人类活动与热机不同，但是与物质一样，人类活动中微观要素间的资源分布和熵增定律也受到相同规律的支配。因此，人类活动和热机一样受到限制。

由于人类活动有很多方面，因此效率的定义也可以是多种多样的。在这里，我们从热力学中类推，选择了最机械性的定义方法，将活动效率定义如下：以活动总时间为分母，以在该活动时间内，目标活动的投入时间为分子，分子分母相除得到的值即为活动效率。用于

目标活动的时间比率，也就是在某项活动中可使用的时间比率，称为活动效率。

如果能按照自己的意志来选择活动，那么活动效率可以提高至100%。但是，如果人类活动遵守热力学定律，那么活动效率就会受到某个上限的限制。熵增定律必须认可活动的“自由”。所谓自由，就是不能将资源全都集中到一项活动上。讽刺的是，认可了自由，就意味着限制了活动时间。

这与热机效率受到熵增定律的限制是同样的道理。正因为原子运动是自由的，热机效率才会受限。

我们在物理学中发现，热机效率上限可以用一个简单的算式来表示，该算式称为“卡诺效率”。卡诺效率的计算方法是“1 减去高温热源温度与低温热源温度之比”。数值越大，卡诺效率越高。

例如，当热机的高温热源是 100 摄氏度（绝对温度为 373 度），低温热源是 0 摄氏度（绝对温度为 273 度）时，卡诺效率就是 1-（273/373）≒0.268，因此效率绝不会超过 26.8%。这是法国军人兼物理学家萨迪·卡诺在研究蒸汽机的效率是否有上限时得出的公式。

如上所述，如果将物质的热力学和人类活动相对应，那么表示热机效率上限的卡诺效率公式也同样适用于人类活动。也就是说，人类活动也存在效率上限。并且令人惊讶的是，卡诺效率公式在数学上也同样成立。我们已经发现的表示人类活动效率极限的公式如下所示：人类活动的效率上限等于 1 减去用于某项活动最活跃的动作值（××次/分）与最稳定的动作值之比。也就是说，该活动中人类活动频带的上限值与下限值限制了效率。

假设写稿时的活动次数是 1 分钟 50～70 次，那么效率极限即“卡诺效率”为 1-　50/70≒0.286，效率极限就是 28.6%。由此我们可以预测在 1 天的活动时间内，写稿时间绝不可能超过 28.6%。

真的是这样吗？我们用前面所说的 9,000 个小时的数据，绘制了所有频带的活动效率图。该图表的绘制方法是，随机选出各种各样的

频带，根据动作的上限值和下限值算出“卡诺效率”的值（横轴），从实际测量数据中，查出在总活动时间中用于该频带的时间比率（纵轴）。由此我们可以证明，人类的活动效率全都分布在以卡诺效率公式为上限的区域中（图 1-4）。

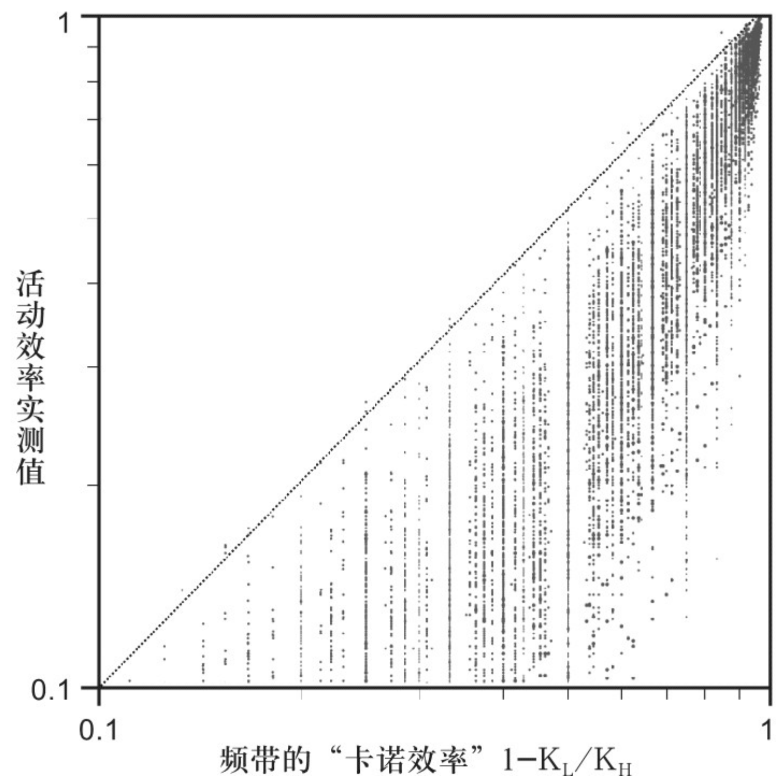


图 1-4 这是证实人类活动效率与热效率受同一公式限制的数据。活动效率受活动频带（频带下限为 K_L，上限为 K_H）的限制。

综上所述，人类活动受到的限制与热机是一样的。

1.17　人的自由与限制

前面我们分析的都是人类行动因自由而受限，下面我们从熵的角度思考一下，人类的自由本身是否受限。

如图 1-1 所示，严格地说，在 230 ~ 300 次/分附近，曲线突然弯折，迅速下降。胳膊不会超速活动，其运动基本以 300 次/分为上限，只有极个别情况才会超过这个上限。专家将这种理论上不会超过的上限称作“Cut Off”。在 U 分布中，与之相似的是 Cut Off 之前的有限区域。

但是，根据人和时间的不同，我们有时也会在除此之外的中间区域内，找到没有完全落在直线上的部分。也就是说，有时多多少少会“偏离”直线。如果人类活动的熵完全实现了最大化，就会完美地呈现出一条直线。因此，“偏离”程度表示了熵值与最大熵之间的“偏离”。由于熵是衡量活动自由度的尺度，因此这种“偏离”显示了自由受限的程度。

通过详细调查这种“偏离”，我们就能将此人在工作和家庭中不自由、受束缚的程度定量化。今后，随着这项研究的不断推进，我们可能会清楚地了解人的压力和精神健康的关系等。在第 2 章、第 3 章中，我们将进一步明确这种行动限制意味着什么。

注 1

指数函数指用$y=a^x$（x是实变量，a是常数）表示的函数。x是 1，2，3……（以一定的间隔变化）时，y就是a，a^2，a^3……。

注 2

关于“累积概率分布与概率密度分布的区别”和“分布名称”，笔者想先在这里做一下补充说明。

我们在讲 U 分布时（例如图 1-1）曾说过，以累计值为纵轴的累积概率分布为指数函数，呈“递降”趋势。实际上，U 分布中，概率密

度分布也是“递降”的指数函数。这是因为将累积概率分布微分后，即可得到概率密度分布，指数函数微分后仍是指数函数。另一方面，吊钟形正态分布图表示的是概率密度分布，比较两者时需要注意这一点。

关于分布名称，可能有的读者心存疑问。前面说过，解释物理现象（物理学）时，人们频繁使用了“递降”分布。微观下热能分配到各个原子，这种现象也呈递降分布。这时，横轴是构成物质的原子的热能，该分布称作“玻尔兹曼分布”。这一分布也在人类行为和社会现象的大数据中频频出现。其横轴多种多样，在这次有关人类行为的测量数据中，横轴是“1 分钟胳膊的活动次数”。本书称之为 U 分布。了解统计分布的函数形式的人，或许会以为这只是单纯的指数分布（用指数函数表示的统计分布）。在第 3 章将说到，从活动频率和密度方面看，U 分布指的是“指数分布”，但是从活动间隔分布方面看，U 分布指的是“幂分布”的实际物理状态，而不只是单纯以函数形式命名的分布。

注 3

我们介绍过，物质中的热能分配给各个分子时，会呈现递降分布——玻尔兹曼分布。构成物质的分子在热能的作用下不断运动，但是每个分子的运动速度（相当于热能）各不相同。大多数分子的热能低，只有少数分子的热能高。制作图表时，如果横轴取分子热能，纵轴取高于横轴热能值的分子数量，图表会呈现递降趋势。但是，其递降方式（坡度）因物质温度的不同而变化。温度高时，热能总量大，分到高热能的原子比例也大，因而递降坡度较小。相反地，温度低时，热能总量小，分到高热能的原子比例也小，因而递降坡度较大。该坡度的倒数称为绝对温度。

① S=klogW——译者注

第 2 章
测量幸福

2.1　能否用科技控制人类的幸福

科技改变了社会，丰富了经济活动，提高了生活水平。德鲁克说，20 世纪体力劳动者的生产力提高了 50 倍。在提高生产力方面，科技功不可没。

我们能够切身感受到，近 30 年来信息技术的发展日新月异。电脑已经人手必备，用电子邮件和手机立刻就能与他人取得联系，建立电子文件夹即可存放文件和发言资料。毋庸置疑，由于这些变化，工作的生产力不断提高。

话虽如此，但科技让我们幸福了吗?

这又是另一回事了。邮箱中积压了大量的邮件，与你有关联的形形色色的人员随便地插话、询问。如果你想认真处理邮件，那么这一整天都可以专心守着邮件。然而，对于一些重要的问题，我们却无暇顾及，人生匆匆而逝。即使是去南方的小岛度假，还是会时刻注意有没有新的邮件和电话——我们的状态是全年无休。

但是，今后科技会不会使我们幸福感倍增呢?

一直以来，科技依据科学知识控制着复杂的体系。例如，保证汽车平稳加速，控制首都圈复杂的列车运行体系等。在科学的控制下，复杂的体系得以维持在理想状态。
朋友圈每日书籍免费分享微信 shufoufou

要控制体系并使其维持在最佳状态，首先必须确定体系的理想状态是什么。关于人类这个体系，人类自己也没有明确的答案。自古以来，宗教和哲学一直在寻求答案。或许这个答案已经在宗教和哲学智慧中了，又或者存在不为人知的秘密。

尽管如此，表现人类理想状态的词语已经被广泛使用了。这就是“幸福”或者“Happiness”。

如果真的能控制人类的幸福，其影响将是巨大的。

然而也有人认为，人类比以往科技控制过的任何体系都要复杂得多，并且幸福因人而异，因国家和文化而异，所以无法一概而论。

但是，我们在上一章中说过，支配万物的物理定律能扩展开来适用于人类。甚至能量、熵和温度的概念也能普遍化，适用于人类行为。

也许我们可以从科学的角度来解读幸福，又或者存在具有统一性的科学。

我们可以控制人类这种复杂的体系吗？可以测量和控制幸福吗？这就是本章提出的问题[1]。

2.2　幸福的心理学——“积极心理学”

实际上，近 10 年关于人类幸福的研究迅速发展。

19 世纪 90 年代之前的心理学，重在治疗和分析有心病的人。例如，在伍迪·艾伦的代表作、奥斯卡金像奖获奖影片《安妮·霍尔》

（1977 年）中，就描述了心理咨询深入影响到纽约知识分子日常生活的情景。

另一方面，大家对健康人的心理状态及其幸福感的研究比较少。不过随着近 10 年的快速发展，这一状况发生了变化，有人开始对“积极心理学”进行研究。

美国加州大学河滨分校的索尼娅·柳博米尔斯基教授是研究幸福（Happiness）的第一人[2]。在此结合教授的《幸福有方法》（*The How of Happiness*）一书，介绍一下幸福研究的概要。首先，幸福是可以测量的吗？积极心理学通过采访和调查，将看不见的人心定量化，以此来研究幸福。让人始料未及的是，仅仅通过回答：

> 1 整体来说，你是个幸福的人吗?
>
> 2 与周围人相比，你觉得自己幸福吗?
>
> （用数字 1～7 来回答。完全不符合是 1，完全符合是 7。）

这种简单的问题，就能将幸福大致转换成定量数值。

这种定量的方法推进了幸福与各种因素的相关性研究，我们逐渐发现，幸福感与我们的直觉并不一致。

首先，幸福受到与生俱来的遗传特性的影响。这是在认真研究双胞胎的过程中发现的。

同卵双胞胎的数据库积累及研究不仅明确了遗传的影响，还推进了有关幸福的研究。同卵双胞胎的遗传特性相同，即拥有相同的 DNA。但是，因为各种各样的原因，双胞胎分别在不同家庭中长大的情况并不少见。当然，在同一家庭中长大的情况也很多。我们要做的就是认真地收集两种情况下的数据。这样一来，我们就能掌握在同一

个家庭中长大的同卵双胞胎和不同家庭中长大的同卵双胞胎的区别。因为这样可以区分开遗传的影响与家庭等成长环境的影响。

通过研究发现，幸福有一半是由遗传决定的。也就是说，既有天生容易幸福的人，也有天生难以幸福的人。

虽然我们都想相信，一切都可以通过努力改变，但是遗传的影响不容否认。

不受遗传影响的另一半幸福将受到后天的影响，并且可以通过努力和环境的变化来改变。这样的话，能改变的部分就比想象的要多。

如果进一步区分后天影响，就会发现一个惊人的事实[3]。

一般来说，当我们找到好的伴侣并结婚，购置新家，拿到很多奖金时，会认为自己的幸福感提升了。但是根据柳博米尔斯基教授的定量研究可知，这些事的影响出人意料地小。

相反地，人际关系恶化、工作失败时，我们会认为自己陷入了不幸，但事实并非如此。不管是好的变化，还是坏的变化，人类都能在短时间内习惯自己周围环境要素的变化，而且所需的适应时间比我们想象的要短得多。

该环境要素包括的内容很广，有人际关系（职场、家庭、恋人及其他）、金钱（广义上的金钱，不仅包括现金，还包括房屋和所有物等广泛的资产）、健康（有无疾病、有无残疾等）。令人惊讶的是，即使我们把这些环境要素全都加起来，对幸福的影响也只占整体的10%。

我在得知这个结果时大受打击，同时怀疑结果是否真是这样。但是，这是经过大量数据的证实和慎重的统计分析后得出的结果。

包括笔者在内，很多人每天都在为改善环境要素而努力，因为我们相信这样做会提升幸福感。然而数据表明，这样做虽然不是毫无用处，却也不会促进幸福感的提升——很大程度上是我们想多了。

那么，剩下的 40% 是什么？这就取决于我们每天行动中的小习惯和选择行动的方法，而且，是否积极采取了行动尤为重要。因为人如果按照自己的意志采取行动的话，就会感到幸福。

如此一来，我们稍微改变一下行动，就能提高幸福感。比如向他人表达谢意，帮助有困难的人。这些事看起来简单，其实会大幅提升幸福感。

采取行动后是否成功并不重要，采取行动这件事本身，就是人的一种幸福。

决定幸福的不是行动成功与否，而是是否积极采取了行动。实际上我们每个人都应该为此感到庆幸，毕竟成功难得。人们以往的看法是，就算早晚有一天会获得成功，但是在那之前必须要忍耐和努力。一直以来，人们都以为在到达幸福的彼岸之前，需要经历一段漫长而不幸的过程，以及长期的忍耐。

如果行动本身就是幸福，那么获得幸福的方法就完全改变了。一种极端的说法是，说不定今天、现在、此时此刻你就可以获得幸福。但是，为此我们必须改变行动。

对人类与科技的关系来说，看法的转变也是一个新的机会。根据上述讨论，假如科技可以为人生带来幸福，那不就意味着科技在帮助人类改变行动吗?

这与以往的科技发挥的作用迥然不同。过去的科技将以往费时费力的作业改用电脑和机器来完成，方便用户使用，并且节省了劳力。更恰当的说法是，在科技的作用下，人类不用自己行动也能顺利完成工作了。

相比以往的科技，以实现幸福为目的的科技发挥的作用是截然相反的。这时科技扮演的是帮助人们主动采取新行动的角色。

提供使人轻松舒适的环境以提升幸福感，属于幸福理论中改善环境要素这一项，其效果顶多占整体的 10%。与之相对，如果人能积极采取行动的话，就能产生 40% 的巨大影响，两者不可相提并论。

但是，困难在于无法指示或命令他人主动采取行动。“主动采取行动”和“指示”是相互矛盾的。我们该怎么解决这一矛盾呢？后面笔者会介绍其中一种答案。

2.3　提升员工的幸福感有利于提高公司收益

另外还要弄清一个疑问：就算我们得到了幸福，会不会仅仅停留在自我满足的层面上呢？

关于这个问题，已经有人用大量的数据做过研究了。

研究表明，幸福的人工作表现好，创造能力强，收入水平高，婚姻成功率高，朋友多，健康长寿。用数据进行定量说明的话，幸福的人工作生产力平均高出37%，创造力高出300%。

重要的是，不是因为工作能力强的人会成功，所以才幸福，而是幸福的人工作能力强。而且，不用非得等到成功来临才能获得幸福，今天稍微采取一些行动，就能提高幸福水平。

当然，幸福感的提升不能保证一定就会成功，但成功的概率肯定会大大提高——享受高效而富有创造性的工作、维持幸福的婚姻生活，健康长寿的概率会提高。

管理领域最权威的杂志《哈佛商业评论》2012年2月号中，有一篇《提升员工的幸福感有利于提高公司收益》的特辑[4]。

迄今为止，公司和员工的关系大多被描述成对立关系。简单说来，人们大多认为公司让员工做很多工作，是公司受益，员工受损；反之则公司受损，员工受益。

当然，为了提高员工满意度、改善工作环境，公司也付出了一定的努力。但是这种努力中包含了很强的目的性，因为如果不将员工的

满意度提高到一定程度，员工的稳定性就会降低，身体和精神健康也会受到损害，这样反而增加了公司的总成本。

比如公司好不容易培养出了有战斗力的员工,他们一旦辞职，就得再花费成本和时间培养新人。也就是说，一直以来提高员工的幸福感被看作一种“必要恶”[①]。

不过，这种现状正在发生变化。人们开始认为，员工幸福地工作既有利于员工本人，也有利于公司。刚刚提到的那篇特辑的编写就体现了这种知识的积累、观点的变化。促进这种变化的，就是以柳博米尔斯基教授为代表的积极心理学朴实的研究成果。

但是，要想用实验具体探讨这一想法，或者估测提高员工幸福感这一对策的效果，以往的心理学方法存在很大的局限。突破这一局限的关键，就是本书介绍的传感器技术。

2.4　传感器可以测出幸福感

我去美国出差换乘航班的时候，在芝加哥奥黑尔国际机场的书店看到收银台前陈列着柳博米尔斯基教授的《幸福有方法》（*The How of Happiness*）。我在飞机上读这本书时发现，可以从科学的角度重新解读幸福。我很希望能和教授共事，于是很快联系到她本人，拜访了她位于洛杉矶郊外里弗赛德市的研究室。

我专门从事人类测量和定量数据解析，而柳博米尔斯基教授专门研究幸福的心理学，此后开始了我们两人之间独特的合作。我为柳博米尔斯基教授介绍了针对人类行为的测量收集技术，而教授也马上认识到这项技术的价值。我想，教授是在这个传感器中感受到了突破性的东西——突破了以往借助调查问卷研究心理学的局限。

柳博米尔斯基教授的研究表明，如果采取对策以提高幸福感，人的行动也会改变。过去很难对人类行为进行测量，有了这个传感器，就可以进行具体验证了。

而且教授认为，一个人的幸福会传递给他周围的人，幸福的范围可以自动扩展开来。但以前没有这样理解人际关系的手段，而借助这种传感器就能实现定量分析了。

不久，我们开始了第一次共同实验。实验对象是某企业的研究开发项目。该项目为了在短期内制造出创新产品，集合了各领域的技术人员。

这个项目的难度很大，一来产品的具体概念尚不明确，二来必须拿出有影响力的创新成果。尽管如此，项目的时间限制却很严格。

项目组长认为，在无法预知未来的情况下，研究成员以很高的积极性共同努力是成功的关键。为此，我们决定测量人类行为，以提供咨询支持，尤其是促进成员积极性的提高，加强他们之间的合作。

作为其中一环，项目组成员参与了柳博米尔斯基教授关于提高幸福感的对策的实验。

参与实验的成员被随机分成了实验群和对照群。两个群组的成员都要写下本周经历的 3 件事。但是，实验群的成员要写本周发生的好事，对照群的成员不能对本周发生的事做出“好”之类的积极评价，而是要保持中立。不论哪个群组，每周只用 10 分钟来写报告，持续进行 5 周。与此同时，我们用问卷的方式调查了成员的幸福感，以及他们对职场的感受和认识，在 5 周后又持续调查了 2 个月。

我们还让这些实验者每天佩戴我们开发的可穿戴式姓名牌传感器（名为“商业显微镜”，英文名 HBM=Hitachi Business Microscope），对其具体的行动变化进行了测量（图 2-1）。

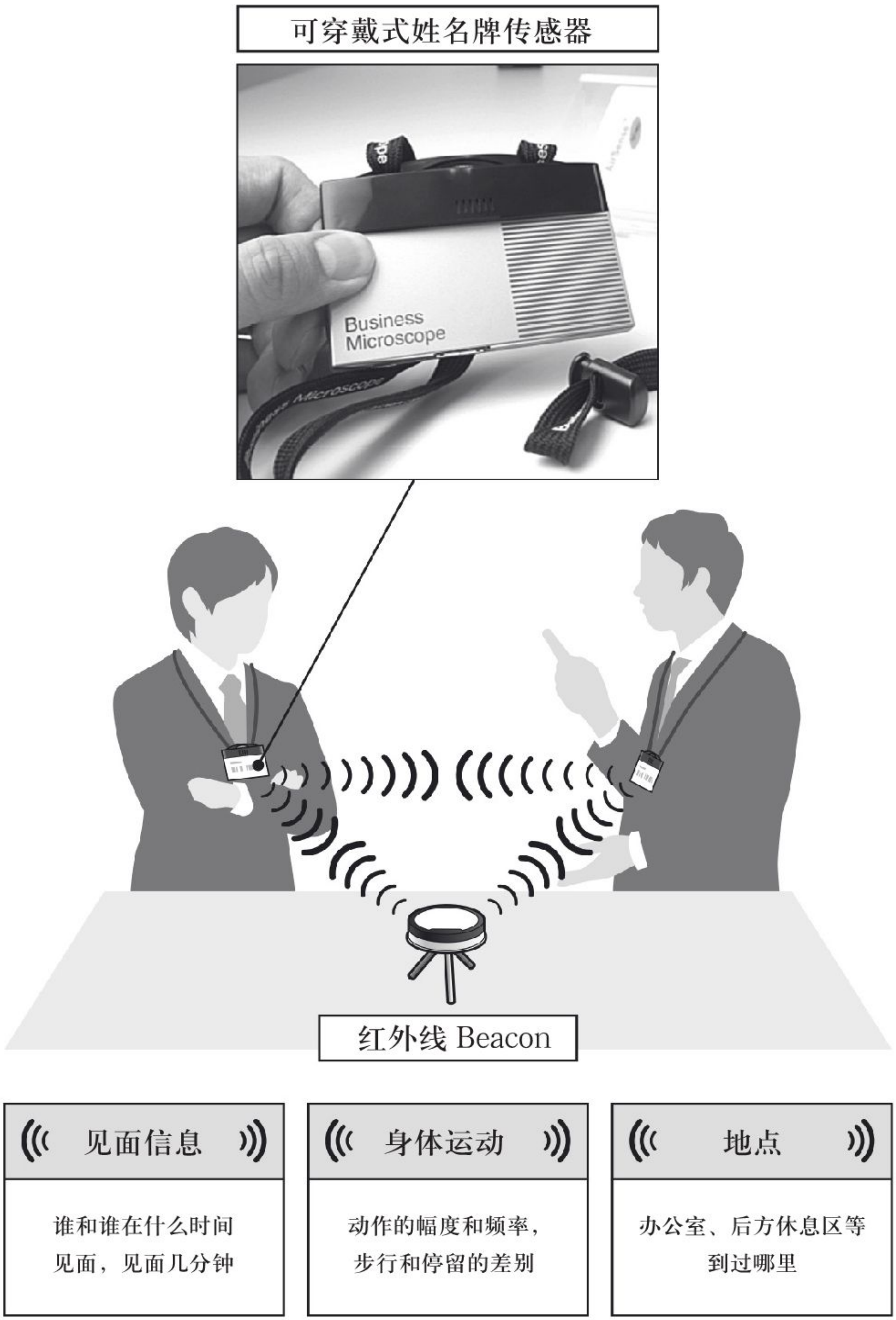

可穿戴式姓名牌传感器
Business
Microscope
红外线 Beacon
见面信息
谁和谁在什么时间见面，见面几分钟
身体运动
动作的幅度和频率，步行和停留的差别
地点
办公室、后方休息区等到过哪里

图 2-1 可穿戴式姓名牌传感器及其功能。可以记录人的身体运动、见面信息和位置信息等。为了记录人的位置，我们要在各地点设置红外线 Beacon。

该可穿戴式传感器是由笔者团队开发、领先世界的产品。在最权威的管理类杂志《哈佛商业评论》2013 年 9 月号上，该传感器被誉为“写入历史的可穿戴式传感器”[5]。

这种可穿戴式传感器的大小与名片一样，形状跟姓名牌相同，可以戴在脖子上。早上上班后，取下插在充电器上的传感器，戴到脖子上，然后正常工作一整天。下班时，将传感器插回充电器后回家。该充电器相当于收集数据的入口。在夜里，传感器的数据将通过网络从该入口传输到数据中心。

我们在该传感器中嵌入了红外线传感器,可以检测出人与人的见面情况。传感器佩戴者能检测出自己面前的佩戴者并进行记录（我们在不同的角度安装了 6 个红外线传感器，不仅能检测出正面的人，还能检测出斜前方或者旁边的人）。在他们开会或者站着说话时，就会记录下那一时间与谁见了面等信息。

除此之外，我们还安装了检测身体晃动和朝向的加速度传感器，以及测量周围音量（可以测出见面时有无对话，但是不会记录对话内容）、温度和亮度的各类传感器。这些时刻变化的物理量连同时间都被记录在姓名牌传感器中，然后被传输到服务器中积累起来。

实验最后的结果清晰明了。经过问卷调查发现，写下好事的成员与中立评价的成员相比，幸福感更高，对组织的归属感更强。

这种内在的心理变化会影响到员工的行动。在本人都没有意识到的情况下，行动就默默发生了变化。

人的活动量会随着时间段变化。借助加速度传感器，我们可以将活动量数据化。人的活动量从早上开始逐渐增加，下午迎来高峰，随后逐渐减少。本次实验中，写下好事的成员从早上开始，活动量的增

速快，高峰时间提前。同时，回家的时间也随之提前。幸福的员工能够很早就活力满满地开始工作，早早地完成工作后回家。

每周只花 10 分钟写下本周的好事，就能产生这么大的效果，着实让人震惊。通过实验我们证实，人的幸福是由微不足道的小事决定的。

还有一个重要的发现是，幸福和身体活动的总量密切相关。也就是说，人内心深处的幸福其实可以通过外部可见的身体活动量来测量。因此，幸福可以通过加速度传感器来测量。

重申一遍，幸福可以用加速度传感器测量出来。

在很多人的印象中，每个人的幸福都是独一无二、各不相同的。

在此我们弄清楚了一个简单而共通的事实，即幸福的人的身体运动多。

当然，根据工作、业务的不同，必要的活动量也会发生变化。但是，我们意外地发现，当一个人感到幸福时，活动频率会增加。

对比工作或者其他条件不同的人，并根据活动量的大小来判断谁更幸福是毫无意义的。但是，更幸福的人活动量的确更大。这和幸福与积极行动密切相关的发现是一致的。

在第 1 章中我们说过，1 天 7 万次的有限活动次数在无意识中限制了时间的利用方法。

我们又发现，人在更幸福的情况下，活动量会增加。人如果采取积极的行动，活动就会增多。如果以第 1 章中热力学与统计力学的知识进行类推，就相当于“活动温度升高”“变热”。而活动的增加，即“变热”成了一种幸福的身体反应。

再者，利用传感器来研究人心，意味着迈出了开拓心理学研究处女地的第一步。

以往的心理学研究通过让被测人回答问题，将无处捕捉的“心”转换为数值。在学术界，这种调查方式称为问卷法。

但是，这种方式很受限制。问卷法类似于抓拍，记录的是回答者在某个时间点的感受，虽然其中包含了重要的信息，但很难捕捉到每个时间点的变化。多次回答同一份问卷的话，会给被测人和调查人双方带来很大负担，可行性不大。并且多次回答同一个问题的话，被测人对问题的理解也会发生变化，进而降低调查的精确度。

因为存在这些限制，很多心理学家都认识到问卷法的局限。由于传感器可以持续测量变化情况，如果能借助传感器测量身体运动，进而观测人心，就可以打破这种局限。因此，人们对用传感器客观测出的人类与组织的海量数据寄予了厚望。

2.5 解读行动中隐藏的符号

在这里，笔者想解释一下将传感器的数据和心理问卷的问题联系起来的重要性。

传感器会客观地记录下每时每刻的测量数据。这些测量数据可以做成波形图，来表示每时每刻的数据变化。比如刚刚在实验中用到的加速度传感器，就可以每隔 1/50 秒记录一次在 x 轴、y 轴、z 轴所表示的空间方向上有多少加速度。

但是，我们无法直接解读出这些大量的、连续不断的数值究竟意味着什么。平均每人每天都会收集到超过 1,000 万个数值系列，这些大量的数字排列一方面以毫秒的速度变化，一方面因年龄增加等影响，以年为单位变化。解读这些复杂的数字排列，是一个很大的课题。

人的行动数据中包含了丰富的信息，可以表现不可见的内心。精神不振时，人连走路都垂着肩膀，举手投足也无精打采。而且对话时的身体活动也能表现出人内心的状态。例如，对话时如果你对对方抱

有好感，身体运动就会很活跃。相反地，如果你不信任对方，身体运动就相对较少。

具体来看，包含幸福在内的人类所有的心理活动和情感，都可以用加速度传感器进行测量，并作为一种行动模式转化成各种各样的符号嵌入数据之中。但是，数据只不过是数字的排列，怎样才能解读其中的意义呢?

柳博米尔斯基教授等心理学家经过多年研究，开发出的检测心理状态的问题发挥了威力。例如下列问题:

> 你是幸福的人吗? 请用数字 1 ~ 7 来回答。
>
> （完全符合是 7，完全不符合是 1）
>
> 或者
>
> 过去 1 周内，你感到过不安和担心吗?
>
> 整体来说，你对工作满意吗?
>
> 你今天发挥出自己的能力了吗?
>
> 你积极解决业务上的问题了吗?
>
> （均用数字 1 ~ 7 回答）

通过这样的问题，我们就能将人的心理状态转换为数值。如此一来，“幸福”“不安、担心”“对工作满意”“发挥出能力”“积极解决业务问题”等心理状态就各自转换成了数值。通过提出这样的问题，能用语言表达的心理状况，就全都可以用数字来表示了。

其中当然也包括误差。不过，如果我们将很多人的回答汇总在一起，就能在“幸福的人”“不安、担心的人”“对工作满意的人”“发挥出能力的人”“积极解决业务问题的人”身上找到共通的倾向，以此来减小误差。将这些数据与调查问卷中得出的大量行动数据结合起来，数据中隐藏的意义便自然会浮出水面。

用传感器测出的行动数据是数字的排列。此后，被测人的身体运动特点也可以用数字来表示。比如将面对面的红外线传感器和加速度传感器的数据结合起来，就可以对在与人见面（从对方身上接收红外线信号）时身体运动是否频繁（1 分钟身体运动几次）进行量化。即可以将“对话时的身体运动量”（每分钟身体运动几次）转化为一定数值。这一数值的大小因人而异。我们将数值大的群体与前文中回答问题的群体进行了对比，结果发现该人群与“积极解决问题的人群”十分一致。即对话时活动频繁是“积极解决问题的人”的共同特点。如果想要积极解决问题，就需要主动地进行对话，这时身体的运动也会活跃起来。

如果通过计算机从大量的行动数据和对问题的回答中搜索出上述对应关系，我们就能理解身体运动中隐藏的意义了。

传感器中记录的大量身体运动和行动数据的排列是一种暗号，我们想要解读这个暗号，却不知道解读暗号的规则。而调查问卷上的回答相当于解读暗号的提示。在有了这些提示后，通过认真调查大量数据，就可以发现解读暗号的规则。

在这个例子中，传感器测出的数字排列中包含的“积极解决问题”这一暗号，实际上可以按照“对话时的活动频率”这一规则来解读。

数据中除了“对话时的身体运动量”以外，还包含其他无数个特点。比如“1 天的活动总量”“每天上班时间的差异”“对话的人数”，等等。我们从数据中抽取了 1 万多个特点，借助计算机调查这些特点与问卷回答的对应关系，以解读人类行为中隐藏的信息。

如果一直推进这项研究，那么将人类行为数据中隐藏的意义全部解读出来，也不是不可能的。人类的遗传信息是大约 10 年前解读出来的，而解读人类行为的重要性不亚于遗传信息。

其中，有关幸福信息的解读价值颇高。如果我们解读出这些信息，就可以使用行动数据测量并控制幸福。这将给个人、商业和行政带来巨大的影响。

2.6　休息时活跃的对话有助于提高生产力

运用定量数据解读人类身体运动中隐藏的意义，并同时控制幸福和生产力。在这方面，我们的努力已初见成效[6]。

我们的实验场所是呼叫中心。所谓呼叫中心，指的是企业的电话窗口。实际上，呼叫中心是最适合定量研究人类行为的尤其重要的场所。

之所以这样说，是因为大多数电话客服人员每天都在重复进行相同的工作——打电话、接订单，因此是否接到了订单、每个电话花费多少时间等大量的定量数据得以保存。这些海量数据就是绝佳的资料，可用于调查人的生产力及对它起决定作用的重要因素。

然而，即使是呼叫中心，以往也并非全程记录了人类的行动。尤其是客服人员离开电话后，在休息区做了什么，以及领导在什么时间下了什么指示等，诸如此类的人类行为并没有被检测出来。

如果借助我们开发的可穿戴式姓名牌传感器技术，就可以网罗上述行动在内的所有数据（称为“人类大数据”）（图 2-1）。我们相信，将大量的业务数据和人类行为数据结合起来，就可以调查出决定人的生产力和幸福的重要因素。

我们在呼叫中心做了实验（照片 1）。这一实验是由 Relia 公司和日立共同实施的。在呼叫中心，需要应对的电话有两种：一种是打电话推销商品和服务（称作 Outbound），第二种是应对顾客打来的咨询电话等（称作 Inbound）。我们选取的实验对象，是推销某项服务的 Outbound 呼叫中心。



照片 1 实验所用呼叫中心（Relia, Inc.②）

我们让实际打电话的接线员及对其监督与协助的管理者等所有相关人员戴上了姓名牌传感器，对其交流情况、不同时间所处的地点、姓名牌的晃动模式等进行了测量。此外，为了掌握他们的个性特点，我们将接线员的技术水平、工作年数等数据联系起来，并进行了问卷调查。我们还收集了“接单率”的数据，即 1 小时内通过电话推销成功的件数。

通过调查，我们发现了一个惊人的事实。整个呼叫中心的接单率每天都不相同。因为有很多接线员一周只工作几天（比如 3 天），所以起初我们认为原因在于接线员每天都会更换。

这样一来，接线员的平均技术水平每天都在变化。一开始我们的预测是，出勤者的平均技术水平高的一天接单率会升高，反之则会降低。

但是，从数据上来看，平均技术水平与接单率的高低没有相关性。虽然相关人员对于技术水平大大影响接单这点深信不疑，但从实际测量结果来看，事实并非如此。

人们一直认为，有些人的性格适合接打电话，而有些人则不适合。但是，经过调查发现，每人的个性特点与接单率之间并无关联。

出人意料的是，影响接单的因素其实是在休息区对话的“活跃度”。休息时，与他人对话的身体运动活跃的话，那天的接单率就高，不活跃的话，那天的接单率就低（图 2-2）。这里所说的对话“活跃度”，是根据接线员戴在脖子上的姓名牌传感器的晃动模式设置的指标（关于指标详情将在后面论述）。如果对话是活跃的，就会产生活跃的身体运动，可以用加速度传感器检测出来。

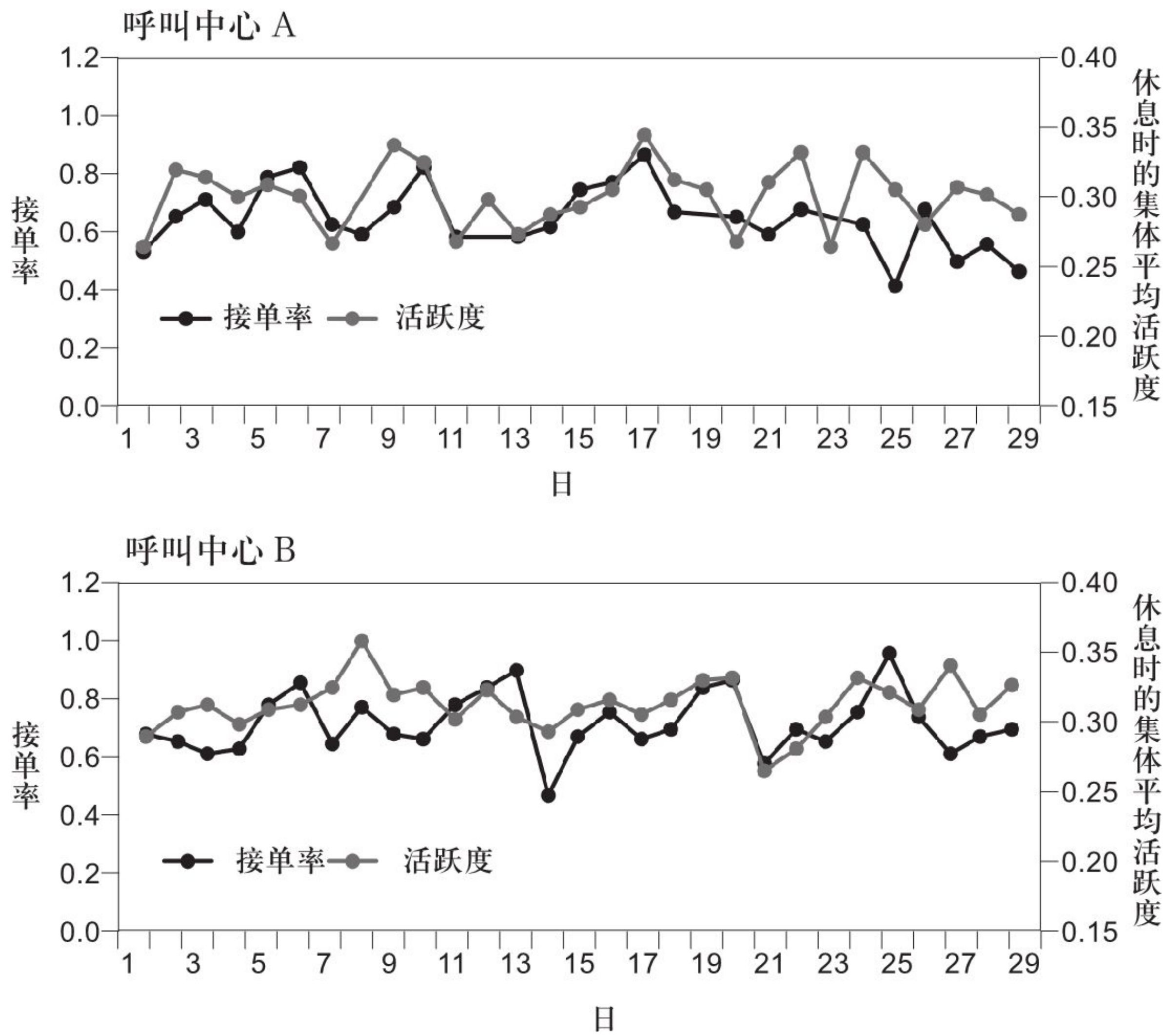


图 2-2 表示每天接单率与休息时活跃度的相关性趋势图

但是，仅凭对话时的活跃度与销售的相关性，我们分辨不出何为因何为果。一方面，我们认为在休息区的活跃对话是因，接单率高是果；另一方面，我们也可以解释为，接单率高的那天，接线员会因工作顺利而情绪高涨，休息时谈话兴致勃勃，身体运动也随之活跃起来。

为了提高休息时间的活跃度，我们让同一年龄层的 4 个人同时休息。结果，他们休息时的活跃度提高了 10% 以上，进而接单率提高了 13%。由此我们发现，通过改变休息时的活跃度，可以改变接单率（图 2-3）。既然因果关系明确了，那么采取极其简单的对策就能大幅度提高生产力。

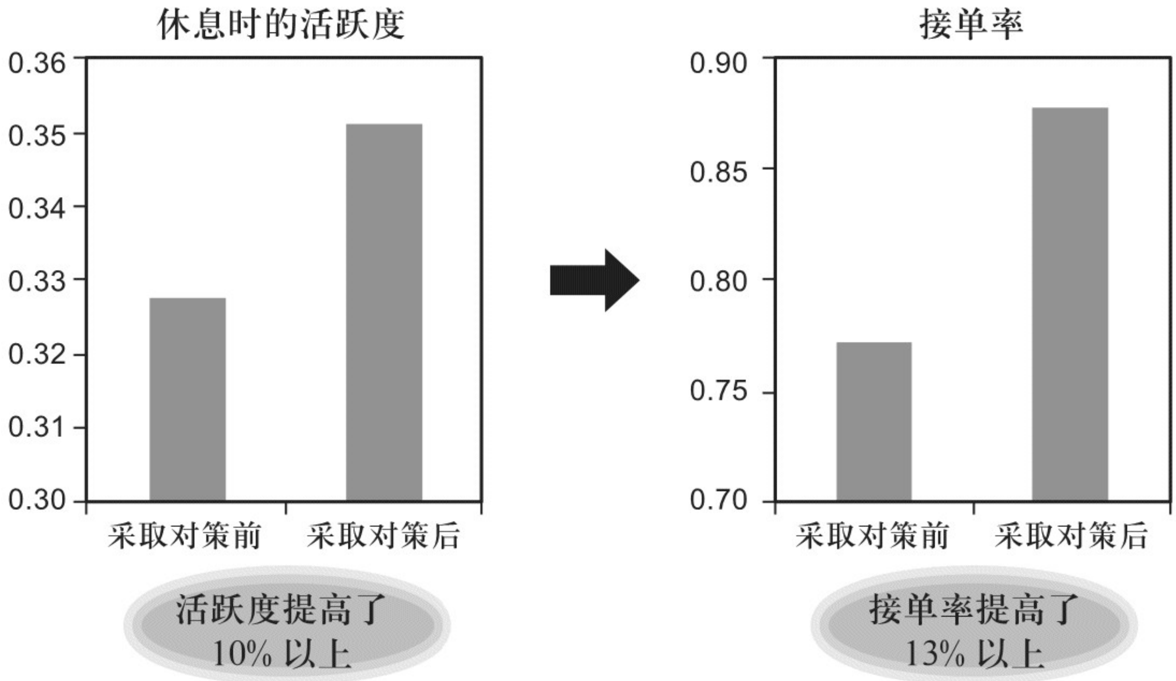


图 2-3 在呼叫中心，通过让同一年龄层的 4 个人一起休息，提高了休息时的活跃度，进而接单率提高了 13%。

但是，这是不是该呼叫中心特有的结果呢？或者说，会不会是这种电话推销服务业务特有的现象呢？难道与该公司的习惯和业务方式没有关系吗？又或者结果仅仅适合顾虑同伴的感受，容易被周围氛围

影响的日本人呢？假如在有着强烈的个人主义倾向的美国，结果是不是就完全不同了呢？对于这些疑问，如果回答不同，对结果的理解也就完全不同。

实际上，我们对于这些疑问已经有了答案。同样是呼叫中心，我们在 Inbound（应对顾客咨询的呼叫中心）也做了类似的实验。该实验是在美国的银行，由麻省理工学院（MIT）的研究小组实施的[7]。

我们在测量接线员的生产力指标，即每个电话的处理时间（从咨询电话打来到结束咨询的时间）时发现，电话处理时间深受接线员在休息时的活跃度的影响。即使把技术、性格、能力等其他影响因素全部加起来，也没有休息时活跃度的影响大。以前接线员们总是错开时间休息，现在尽量让他们一起休息，通过这一对策，生产力最大提高了 20%。而整个银行通过采取该对策，减少了 12 亿日元的成本。

重点在于，接单率与 Outbound 和 Inbound 的业务、美国和日本的差异毫无关系。虽然 Outbound 和 Inbound 同为呼叫中心的业务，但性质并不相同。而且日本和美国的企业在业务规范、职场流程以及常识上都存在很大不同。尽管如此，结果却是相同的。

由于呼叫中心保存了定量数据，因此可以进行验证，但其实这并非呼叫中心特有的现象，而是人们工作时的普遍现象，可以适用于呼叫中心之外的其他业务。

用一句话概括结论，就是在活跃的职场，员工的生产力会提高，而在不活跃的职场，员工的生产力会降低。这是一个普遍的倾向。再者，如果用加速度传感器测量集体员工的身体运动情况，就可以实现职场活跃度的定量化，从而可以在各种各样的产业中，确认活跃度与生产力的关系。

有意思的是，这个实验的接线业务中，个人发挥的成分居多，集体发挥的要素较少。即使是在偏于个人发挥的业务中，职场活跃度这种群体性因素也会对生产力和成本产生极大的影响。人们迄今没有认

识到这种相关性。因为按正常思路来思考的话，这两者之间应该没有关系。

追究起来，两者之间为何会产生联系呢？

2.7 身体运动会传染，幸福也会

从表面上看，员工的活跃度很抽象，如何把抽象的活跃度定量化呢？关键就在第 1 章论述的身体活动。

借助加速度传感器，我们可以测出员工每分钟的活动次数。如果每分钟的活动次数高于基准值（按照以往的测量经验设定）则算作活跃，低于基准值则算作不活跃。按照这个判定标准，我们将处于活跃状态的时间占规定时间的比率称作活跃度。假如 1 个小时中活跃的时间有 30 分钟，那么活跃度就是 0.5。此外，对一个群体而言，取全员活跃度的平均值，就可以得到群体活跃度的数值。这样一来，我们就将员工整体的活力程度或活跃程度定量化了。

用活跃度这个指标来理解人类和社会，表面上看可能有些刻板。但是，我们居然可以借助这个看似刻板的群体活跃度，判定员工在呼叫中心的电话业务中的表现（与业绩成正比）。

更为重要的一点是，身体运动的活跃度会从一个人传染给另一个人。周围人的身体运动活跃的话，自己也容易活跃起来；周围人的身体运动停滞的话，自己也会停滞[8]。这一点已经通过大规模的数据测量与分析得到了证实。

哈欠会从一个人传染给另一个人，这是我们经常经历的事情。而身体运动的活跃度也会传染，并产生一系列的连锁反应。也就是说，活跃的身体运动和停滞的身体运动也在互相传染。

这和物理学中的磁性原理十分相似。具体来说，磁铁内部微小的磁铁（称为 spin）相互统一方向，从而形成了 N 极和 S 极。由于相邻的 spin 倾向于同一个朝向，因此产生了统一的方向，即形成了磁铁。物理学中，将这种群体自发形成的现象称为“协作现象”。

在人类群体中，人们经常通过开会或站着谈话等方式，与周围的人相互影响，这时发生的活跃度传染的现象也属于协作现象。我们已经证实，可以通过解释磁铁特性的物理模型（称为“伊辛模型”），对活跃度传染的现象加以定量说明。在多人聚集的场所，人的身体运动会产生群体性协作现象，人与人之间的相互影响，决定了其活跃度。

我们往往简单地认为，自己的身体由自己（或者自己的大脑）决定。但是这并不正确，我们深受周围人的影响，同时也影响着周围的人。

也就是说，群体内的身体活动中存在一种连锁反应。一个人的身体活动一旦活跃起来，就会产生一系列的连锁反应，使得周围人的身体也随之活跃起来。不知不觉地，我们就成了这种连锁反应的一部分。

而且，身体运动的活跃度与幸福的相关性在此具有重大意义。因为身体运动会传染，这就意味着幸福也会从一个人传染给另一个人（前面说过，幸福与个人活动量的增加息息相关）。如果将我们主观感受到的幸福视作随着群体运动的活跃而产生的感觉（多半是后来产生的感觉和意识），那么实验事实就符合逻辑，可以理解了。

如果我们承认了这一点，那么幸福就成了一种群体现象。与其说幸福是从个人内部产生的，不如说幸福是在群体中，因人与人之间的相互作用产生的。并且如果群体中产生了幸福，那么企业的业绩和生产力就会提高。

我们已经证实，即使是呼叫中心这种看似个人发挥较多的接线业务，群体的力量也会对结果产生极大的影响。在身体运动会产生活跃连锁反应的现场（一般的说法是有活力），生产力就会提高；反之，

在身体运动难以产生连锁反应的现场，接线员身体运动的开关就会闭合，生产力随之降低。

在活跃的身体运动的作用下，呼叫中心的生产力提高了 10%～20%。但是，很明显这个数值是最低值——因身体运动的连锁反应而提高的活跃度在此对生产力的影响是最小的。之所以这样说，是因为在比呼叫中心更注重群体性的业务中，群体的力量将发挥更大的作用。

为了发挥这一效果，必须推动人们采取行动。指示别人去主动采取行动，在表达上就自相矛盾，但是我们可以创造一个有利于主动活动身体的环境。

其中，休息时间或者午休的环境尤为重要，因为休息时间是如何度过的，将影响到之后的业务。经实验证实，如果在休息时间能跟别人愉快地聊天，进而使身体运动产生活跃的连锁反应，之后的业务生产力就会提高。实验还证实，如果领导和管理者主动跟员工打招呼，全体员工的身体活跃度也会提高。

在这里我想起了一件事。以前，日立每年秋天都会在各地的事务所举办员工运动会。为此日立投入了大量的时间、精力和经费。当时我还年轻，印象中运动会前的 1 个月就没怎么工作。工作期间，我们还在开“扳倒旗杆”[3]和“拔河”的作战会议。有时候，有人因为工作的关系提出无法参加练习和作战会议，前辈就会训斥：“你分不清运动会和工作孰轻孰重吗？”前辈希望对方回答的是“当然是运动会更重要”。

1990 年前后是日本的泡沫经济时代，社会上充斥着浓厚的个人主义氛围，很多人主张“时代已经变了”（应该废除运动会），自此运动会便逐渐消失了。

现在回想起来，当时运动会作为公司的一项大型活动，大幅提高了现场的活跃度，促进了生产力的飞速提升。“运动会和工作孰轻孰重”的潜台词是“运动会更重要”。的确，哪怕是削减工作时间，也应该举办运动会——这是前人的直觉。如前文所述，现在我们已经有了证

明运动会的作用的科学证据。如果当时就有这样的证据，想必运动会就不会被废除了。

2.8 身体运动活跃的职场的优点

前面我们得出了一个结论：身体运动的连锁反应会对商业产生极大的影响。事实上，这个结论不仅适用于呼叫中心，在其他很多地方都适用。

迄今为止，我们用传感器在很多公司测量了员工的身体运动，并以员工为对象进行了问卷调查。

其中包括来自 11 个组织的 630 名员工的数据组，这些员工从事软件开发、研究开发、设备设计、管理等休息时间不确定的业务。分析这些数据后发现，参与问卷调查的员工的平均压力水平与加速度传感器测出的群体身体运动的活跃度（员工的平均活跃度）有很大关联。职场平均压力水平高，员工在群体中的平均活跃度就低；反之，职场的平均压力水平低，员工在群体中的平均活跃度就高。

另一方面，从问卷调查得出的幸福水平（主观幸福度）来看，在群体平均活跃度高的职场，员工的幸福水平高；反之，在不活跃的职场，幸福水平比较低。

前面已经说过，幸福水平高的话，业务生产力会提高 37%，创造性会提高 300%。结合这点来考虑，我们发现，能否在职场中通过身体运动的连锁反应来提高活跃度，或者创造一个可以提高活跃度的职场环境，直接关系到公司的业绩。

还有另外一组数据，调查对象是能源、信息、电子、材料等多个领域的技术人员。虽然他们同为技术人员，工作内容却大不相同。例如，信息领域的技术人员一整天都对着电脑开发软件；能源领域的技

术人员会在现场做实验，以启动大型发电机；电子领域的技术人员或是在洁净室试制产品，或是在实验室测试电路。他们各自的业务和领域不同，沟通量也截然不同。

但是，对众多领域的技术人员来说，对话时身体运动的活跃度与问卷中“你在积极解决问题并努力创新了吗”的答案有很大的相关性。在身体运动活跃的对话现场，人们会积极解决问题，并努力创新。而在对话时身体运动不活跃（身体运动少）的现场，人们不会积极解决问题或努力创新。积极发现问题并别出心裁地解决问题的技术人员和不积极解决问题的技术人员相比，工作成果和生产力会有天壤之别。

重要的是，对话时的活跃度不是主观的产物，而是根据传感器的测量数据推算出来的、有明确定义的客观指标。前面说过，所谓活跃度，指的是身体运动超过基准值的时间占规定时间的比率。对话时活跃度的大小与问卷中“积极解决问题并努力创新了吗”的回答数量息息相关。即使从个人角度来看，对话时身体活动频繁的人，更倾向于积极解决问题并努力创新。相反地，对话时身体活动少的人，不太倾向于积极解决问题或努力创新（在这里，我们比较的不是同一个人在一段时间内的身体运动，而是不同的人之间的身体运动。而关于幸福感，我们比较的是同一个人在一段时间内的身体运动。还要注意的一点是，我们测量对话的活跃度，并不是看每个人 1 天的活跃度）。对话时长因人而异，这与“积极解决问题并努力创新”一题的回答数没有关系。因为对话量的多少因工作种类而异，单看对话量无法判断与此问题的关系。

从群体的角度来看，对话时身体运动活跃、群体平均活跃度高的职场，可以让人很容易地进入积极行动的状态。这样一来就形成了率先发现问题，别出心裁地解决问题的职场。

综上所述，我们可以从大量的数据中发现一个简单的结论。即人的身体运动会诱导周围人的身体运动，在这种连锁反应的作用下，会产生群体性身体运动。如此一来，人们就进入了积极行动的状态，员工的幸福感和生产力都得以提升。

在呼叫中心这种以个人作业为中心的业务中，积极的行动可以促使生产力提高 10%～20%，在更需要团队作业的业务中，生产力有望提高 37% 以上。而在那些更加要求创造性的业务中，甚至可能提高 300%。

在 Gap 公司，我们以 1,000 多万名员工为对象，调查了员工行动积极带来的影响。员工行动积极的公司和行动消极的公司相比，平均每股的利润率差异高达 18%。因此，其因果关系如下所示：

员工身体运动的连锁反应，促进活跃度的提升

↓

员工的幸福感和满意度提高

↓

高生产力与高收益

2.9　打造活力职场是一项重要经营项目

身体运动的活跃度是一个可测值，平常的说法是，将现场的活跃程度（或者生气和活力）转化成了数值。如果被问到“是否充满活力的职场比较好”，想必很少会有经营者给出否定的回答。

但是，我们要认清一个事实：一直以来，打造活力职场在企业经营中并不处于优先地位，经营者也从未在这上面投入时间和资金。其原因在于，过去人们没有明确认识到职场的活力与公司收益具有相关性。

在企业经营中，经营者为了获取收益而决定经营方针，并为此实施具体的经营对策。这些做法都需要逻辑解释。一直以来，虽然人们觉得组织最好是充满活力的，却没有一种逻辑将活力与收益联系起

来。因此，活力不处于优先地位。但是，我们的一系列研究为此打开了突破口。

如果现场的活跃度提高，生产力和收益也随之提高的话，对经营者来说，在提高现场活跃度方面的投资便有可能成为最有效的投资方式。这样一来，我们就开辟了一条新的道路——以较低的投资来提高收益。

相反，如果经营策略具有副作用，会降低现场活跃度，那么就会使收益相应地减少。迄今为止，人们还没有具体考虑过这一影响。

2.10　我们也要考虑 IT 会降低生产力

现场的活跃度与生产力及成本直接相关，实际上也会大大影响企业的信息技术（IT）。

近 20 年来，在 IT 系统的作用下，企业的业务环境焕然一新。为了导入 IT 系统，企业投入了巨额资金，并在此后的系统运营和维护上不断投资。

人们为了提高业务效率和生产力，设计出了 IT 系统。因此，在开发业务用 IT 的过程中，人们一直都是通过分析业务流程（程序），并根据该流程来设计 IT 系统的。

然而，在以往的 IT 设计中，人们完全没有考虑到现场活跃度的重要性。

导入 IT 之后，表面上看效率提高了，但其实很多情况下，IT 的导入会排除群体身体运动的连锁反应所需的要素，进而导致能够促进活跃度连锁式提高的体系瓦解，结果反而降低了生产力和创造性。IT 本应是提高生产力的工具，有时却成了降低生产力的原因。

比如，过度依赖邮件会剥夺人们原本应该直接面对面、通过身体运动进行交流的机会。以前，下属会请领导在纸制文件上盖章，并借此机会通过身体运动来了解领导的想法和工作的优先顺序。而现在，由于领导的审批程序已经 IT 化，下属也就失去了这个机会。我们忽视了导入 IT 系统的负面影响，而这是否造成了业务生产力的下降呢?

从更宏观的角度来看，这是否是造成日本近 20 年来优势地位急剧下降的原因之一呢？20 世纪 80 年代之前，日本企业通过活跃职场推动了员工的积极行动，进而促进了经济的迅速发展。但是后来企业的生产力和创造性都下降了，而这时正好是企业导入 IT 系统的时候。当然，美国等其他发达国家基本也在这个时候导入了 IT 系统，但不同之处在于，日本对职场活跃性的依赖度更高。因此，职场活跃度的下降对日本企业的生产力和创造性的影响更大。

现在，是时候依据有关人类的科学知识，重新构建 IT 系统和经营体系了。新的 IT 系统必须能够推动员工的积极行动。

2.11 通过幸福科技创造幸福指标

如果任何人都能随时随地、精准测量并控制自己的幸福，那么不仅是工作，社会的方方面面都将发生变化。

若家人共享幸福的数据，家人之间就能互相帮助。毕竟人生不是一帆风顺的，有时我们也会面临危机。这时，如果有可靠的家人不计得失地支持我们，对我们将是巨大的帮助。借助传感器技术，即使与家人天各一方，也能彼此共享人生的变化，互相扶持，甚至是与年迈的父母分开生活了，也能互相感知到对方的情况。

在行政领域，人们一直将 GDP 作为衡量国民生活质量的常用指标。但是，诸多研究调查表明，GDP 一旦超过最低水准，就与国民幸

福度无关了[9]。因此，英国、法国、奥地利、不丹和日本等国都表示，要创造一个超越 GDP 的新指标。

如果我们能使用传感器技术测量幸福，那么国家领导人就可以实时监测国家的规章政策是否与国民、社区的幸福度挂钩。这将从根本上改变政治和行政流程，或许从此就可以采取一种前所未有、灵活崭新的方式治理国家了。

① 由于组织运作或社会生活的需要不得已而为之的事情。——译者注

② Relia, Inc.是日本一家大型业务流程外包（Business Process Outsourcing）服务公司。——编者注

③ 两支队伍各有一根旗杆，在保证自己一方旗杆不倒的同时推倒对方的旗杆即为胜利。——译者注

第 3 章
求“人类行为的方程式”

3.1　人类行为中存在方程式吗

在第 1 章中，我们根据身体活动科学定量地分析了人类行为，证实了人类行为完美遵循 U 分布统计规律。虽然从表面看来，人的性格、想法、成长环境等各有不同，但实际上，人们身体运动的统计分布曲线呈现同一形状。这是一个惊人的事实，也是一个巨大的发现。接着在第 2 章中，我们介绍了身体运动是人类的终极目标——幸福的来源。

在本章，我们将进一步分析并证明：从统计规律中可以导出人类行为的方程式。

回顾科学的历史我们会发现，科学中方程式的出现是人类的一大转机。在距今 400 年前的 17 世纪，正处于近代科学的黎明之前。当时，日心说和地心说两种观点相互对立，天文学家和占星学家被混为一谈，女巫审判案件的情况屡屡发生。

此时，丹麦天文学家、占星学家第谷·布拉赫收集了大量有关天体运动的数据，当时担任他助手的是德国的约翰尼斯·开普勒。在布拉赫突然去世后，开普勒认真研究了布拉赫留下的天体测量数据。他发现

行星并不是沿着过去认为的圆形轨道运转，而是沿着椭圆形轨道运转，这就是所谓的开普勒定律。

在从大量的数据中发现统计规律这一层面上，第 1 章介绍的 U 分布的发现或许就类似于开普勒定律的发现。实际上，我们研究小组收集了超过 100 万天（同一个人不同的一天也算作 1 天）的详细人类行为数据，这些数据大显神威，与开普勒定律的情况很相似。

继开普勒定律之后，方程式的出现代表了真正科学时代的开始。牛顿发现了物体运动的方程式，借助该方程式，我们可以统一理解并预测从天体到苹果的各种物体运动。从这个意义上来说，方程式就是推动后来科学发展的“魔法杖”。

牛顿创造的方程式这一“魔法杖”超越了物体运动的范畴，陆续扩展到了其他领域——电磁现象中的麦克斯韦方程，流体现象中的纳维–斯托克斯方程，热现象中的玻尔兹曼方程，原子运动中的薛定谔方程……可以说，方程式的发展史就是这 300 年来科学的进步史。自此，人类确立了所有自然现象都可以用方程式来理解的科学体系。毫不夸张地说，现在理工科大学的教育，就是关于这些方程式及其应用的教育。

然而，目前方程式仅在物质现象和自然现象中成立。人们普遍认为，在社会现象和人类行为中不存在类似的方程式。之所以这样认为，是因为人们觉得人类和社会太过复杂，无法用方程式表示出来。

在本章，笔者想和大家一起跨越这个难关。

3.2　方程式究竟是什么

在寻求人类行为的方程式之前，我们先梳理一下方程式究竟是什么。

方程式的特点在于，用统一而普遍的定律来总结多种多样、各不相同的现象。在人类社会中，每天都有各种各样的人聚在一起，发生着形形色色的事情。这时，人们往往会认为，“在如此多样的现象中，怎么可能有统一的定律”。

“或多样，或统一”，这话听起来好像是说，只能有一方正确，另一方错误。但是回到300年前的话，月亮和苹果分别属于天体和水果这两种完全不同的种类，而从中发现统一定律的就是牛顿。月亮和苹果跨越了种类的差异，遵守着相同的运动定律。性质不同与遵循统一定律并不矛盾。最先察觉这一点的是牛顿，此后科学家们将这一原理彻底应用到了所有事物上。

为了合理解释现象的多样性和规律性，方程式要具备某个特点，即表示在时间轴（或空间轴）上发生了怎样的急剧变化（状态量的变化用“坡度”或“倾斜度”表示。在数学上称为“微分”）。我们关注的不是状态本身，而是状态量的变化（倾斜度）。

例如，我们手上拿着苹果，在放手的瞬间，苹果的状态即其位置和速度就开始变化。速度以重力加速度g开始加速，然后不断向下加速。牛顿方程式证实了一个原理：物体每时每刻都在从现在的状态转变为下一个瞬间的状态（Generate）。从这个意义上看，“Generator”就是创造出每时每刻变化的体系。

但是，该物体有多大、一开始在哪里、往哪个方向移动了，以及什么时候会施加什么力等，并不受方程式的限制。就像前面举的月亮和苹果的例子一样，方程式的对象可以是多种多样的。通过改变方程式的参数（常数和限制条件），就可使其适用于各种情况和物体。这样一来，遵循统一的方程式（即Generator）和多种多样的现实之间就不矛盾了。科学之所以能适用于宇宙万物，就得益于方程式的这个特点。

在这里，我们可以用“微分”一词来表示从现在的状态转变成下一个瞬间的状态（Generator）。为了科学地理解世界，基于方程式的上述特点，我们必须在方程式中使用微分。牛顿和莱布尼茨之所以创造

了这种新的数学工具——微分，是因为迫切需要使用数学方程式，合理且科学地同时表示现实的多样性和统一性。

如果在对人类和社会的研究中我们也能利用好方程式的这个特点的话，就可能在不违背各种社会现实的情况下，找到统一的定律。下面就让我们来具体看一下。

3.3　与人的再次见面遵循普遍定律

为了在人类行为中找出方程式，我们必须先找到创造出每时每刻变化的规律。也就是找到我们行动、人生的 Generator（定律）。介绍这些需要一些数学工具,笔者在介绍时尽量不使用数学公式。

我们已经说过，创造出每时每刻变化的体系是 Generator，在定量地表示 Generator 时，微分是一个重要的工具。

但是，仅仅靠简单的微分无法捕捉人类和社会，因为人类和社会中存在很多不连续的变化，而微分原本是分析连续变化的物体运动的工具，表达的是变化的坡度（倾斜度）。如果用微分来表示不连续的变化，变化的坡度（倾斜度）就会被无限放大。

例如，身体是否活动、是否与领导见面等，这些状态都是不连续且分散的。动或者不动，见或者不见，是非此即彼的关系。人类就在这种不连续且分散的状态之间游走。

下面举一个具体的例子。试想一下，当我们与人见面时，是否可以找到创造变化的 Generator 呢？我们与人见面，分别，再见面。见面的频率各不相同。也许和某个人每天都见面，和另一个人每周见 1 次，和有些人的见面时间并不确定。

我们将上一次见面与下一次见面之间的时间称为“见面间隔”，以此作为衡量见面现象变化的定量指标。例如，假设你和你的领导藤田

科长一起吃午饭，下午 1 点分开。接着你又在下午 3 点的会议上见到了藤田科长。这时的见面间隔就是 1 点到 3 点的 2 个小时。

我们来思考一下从没见面状态变成见面状态，这种分散活动发生的概率。这一概率有可能是每秒 10%，也有可能是每秒 30%。我们将此概率看作 Generator（这涉及专业知识，大家可以跳过不看。其数学定义是，分别之后在时间t之前没有再次见面的条件下，在接下来极短的时间Δt内发生活动的概率密度）。也就是说，它表示的不是活动的不连续变化，而是概率的变化，在这里也会使用微分。

如果见面是以一定概率随机发生的，那么它在统计学中就会遵循泊松分布（类似于第 1 章中提到的正态分布）。你站在路边等出租车，直到遇到空车前的这段时间遵循的就是泊松分布。路上平均有多少空出租车是一个确定的统计值，只不过运气好的话立刻就能遇到，运气差的话则要等很长时间。经过多次试验，可以得到平均等待时间——这就是泊松分布。实际上，遇到出租车这一活动的 Generator 相当于平均等待时间τ（希腊字母 tau）内遇到 1 次出租车的概率，可以用方程式来表示（后面论述）。

与人见面，比如与藤田科长见面的概率是怎样的呢？它遵循的是平均 1 小时见 1 次的泊松分布吗？

利用我们开发的可穿戴式姓名牌传感器[1]，就可以将人们见面间隔的实际情况转换为定量数值，从而验证见面间隔是否遵循了泊松分布。

前面说过，该传感器的大小和名片一样，形状与姓名牌相同，里面嵌入了红外线传感器，可以测出传感器佩戴者之间的见面情况。而且里面还安装了加速度传感器等，既可以检测出自己面前的传感器佩戴者，又可以检测出身体的晃动和朝向。这些有关物理量的数据组每时每刻都会被记录在姓名牌传感器中，然后传输至服务器并积累起来。

我们使用该传感器收集了大量现实社会中的见面数据后发现，见面概率与时间的关系并没有呈现泊松分布。迄今为止，我们测量了大量人与人之间见面的数据，共计超过 100 万天。从经营者到新员工，从技术员到营业员，包含了各种各样的人之间见面或者不见面的数据。

我们根据大量的数据，在分析了时而与人见面、时而独自一人的这一变化后发现，距上一次见面之后，时间过得越久，再次见面的概率就越低。将距离上一次见面的时间用T表示，再次见面的概率就与$1/T$成正比。

假设你和藤田科长见面之后已经过了 1 个小时。我们将再次见面概率用P表示，那么 2 小时后的见面概率就是$P/2$，3 小时后的见面概率就是$P/3$。这个规律既适用于公司领导，也适用于新人；既适用于营业员，也适用于研究人员。

用一句话来说，我们已经证实，距离上一次见面的时间越久，就越难再见（见面概率下降）。而且这一规律完美地遵循了反比的定律，我们将其称为“$1/T$定律”。

各种领域的人都像被无形之手操纵了一样，遵循着 $1/T$定律。人们是按照统一的定律来行动的。

这也就意味着，我们发现了引发见面活动的 Generator。

3.4　以见面概率为基准思考，则时间的流逝各不相同

“去者日以疏。”

古人用这句话来表示，即便是亲密之人，若长久不见，也会渐行渐远。从另一个角度来看，也可以解释为，我们与远去之人的时间流

逝是不同的。

实际上，结合刚刚经过大量数据分析得出的结果，这句话可以作为一个定量法则，即时间的流逝是时快时慢的。

举个例子，假设你需要在各个工作阶段见到藤田科长，向他进行汇报、请求审批，否则工作就推进不下去。在这种情况下，与藤田科长的见面，就在你的工作中发挥了钟表的作用。也就是说，你与藤田科长的见面概率一旦下降，工作上的时间就会流逝得很慢（工作毫无进展）。

反过来，我们以工作时间流逝为基准，重新审视一下物理学意义上的时间流逝。与藤田科长见面后，物理时间流逝，见面概率降低，工作时间流逝变慢，工作停滞不前。为此，我们要以工作时间为基准，重新把握物理时间。这样一来，钟表看起来会走得很快（也就是说，工作停滞不前，物理时间却马不停蹄地向前走）。两天后你会感觉钟表转动的速度比前一天快两倍。如果 4 天都没有见到藤田科长，就会感觉钟表快了 4 倍。也就是说，时间流逝的速度是不同的，见面间隔越长，就感觉时间过得越快。

如果我们对人类和社会科学反复进行定量分析，就会重新发现自古以来广为人知的智慧。但是，科学定量数据至关重要，它与单纯地引用古语有本质区别。古人所说的“去者日以疏”，或许只是他当时抒发的主观感受，因此我们可以说“我不这么认为”“现在时代已经变了”。但是在有了数据后，一切都变得清晰了。

3.5　1/*T*定律也适用于回邮件等其他行为

实际上，1/*T*定律不仅适用于见面这一种情况。

例如，美国东北大学的艾伯特–拉斯洛·巴拉巴西教授收集了大量的数据，调查从收到电子邮件到进行回复的时间，并进行了解析[2]。巴拉巴西教授没有从决定时间概率变化的 Generator 的角度进行解读，而是集中精力分析了一种叫幂律分布的统计分布曲线。

我们从 Generator 的角度重新分析了一下这些数据后发现，距离收到电子邮件的时间越久，回复的概率就越低。如果回复之前的那段时间用T表示，那么回复概率与T成反比（与 $1/T$成正比）。也就是说，从收到电子邮件到进行回复的这段时间遵循 $1/T$定律，这段时间越久，回复的概率就越低。

此外，东京大学的中村亨先生等人还调查了人们在日常生活中的安静状态能持续多久（身体活动少的状态）[3]。安静状态会因起身、被人搭话而打断。我们借助加速度传感器检测人何时会从安静状态转为活动状态。该研究没有在 Generator 的角度上进行解析，而是集中精力对活动的发生频率进行了统计分析。

我们从 Generator 的角度重新分析了这些数据后发现，假设安静状态的持续时间为T，那么转为活动状态的概率就是 $1/T$。从持续了 1 小时的安静状态转为活动状态的概率是 1 的话，从持续了 2 小时的安静状态转为活动状态的概率就是 1/2。由此可知，$1/T$定律在这里也成立。安静状态持续得越久，越难转为活动状态。

更重要的一点是，中村先生的数据包括对精神健康者和精神抑郁者的比较数据。从 Generator 的角度上分析这一数据会发现，精神健康者和精神抑郁者都会遵从 $1/T$定律，从安静状态转变为活动状态。但是，就转变概率而言，精神健康者比精神抑郁者大约高出 20%。

也就是说，如果我们测出由安静状态转变为活动状态的概率，就可能掌握人在受压力影响时的变化情况。该转变概率可以用可穿戴式传感器测量得出。这样一来，我们就可以对自己承受的压力大小进行简单的确认了。

我们还有一个重要发现。在与东京工业大学的三宅美博教授的共同研究中，伴随肢体活动的人类行为的持续时间一般遵从 1/*T*定律（图 3-1）[4]。我们在分析了可穿戴式传感器测出的有关人类行为的大量记录后发现，一旦活动开始，时间越久，停下的概率就越小。用*T*表示活动从开始到停止的时间，活动中断概率完全与 1/*T*成正比。*T*越大，概率越小。当然，1/*T*定律是有极限的——持续时间基本为 20 分钟到 100 分钟。到底能持续到什么时候，还要看当时的环境条件。

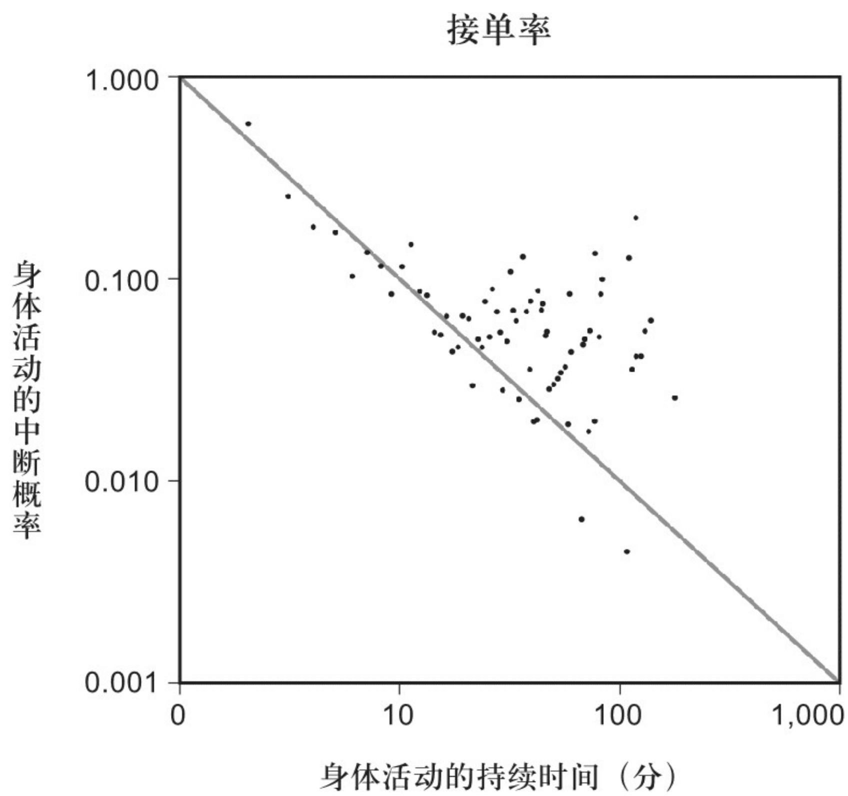


图 3-1 身体活动的持续时间为*T*，*T*越大，身体活动就越容易持续下去（中断概率下降）。至少在大约 10 分钟以内，中断概率与 1/*T*成正比。请注意在这张图中，纵轴和横轴的数字均为对数。

3.6　行动持续越久越停不下来

如前所述，从上次见面到再次见面的时间间隔，从收到电子邮件到回复邮件的时间，从安静状态转为活动状态的时间，伴随肢体活动的人类行为的持续时间，这 4 种活动均遵循 $1/T$定律。这表明，Generator 在各种人类行为中均发挥着基本作用。

关于该定律，通俗地说就是持续得越久越停不下来。不管是不与某个人见面、不回电子邮件，还是安安静静休息的状态、伴随肢体活动的行动等，都具备持续得越久就越停不下来的性质。

如果想要分别说明这 4 种行动中共通的 $1/T$定律，那可以说的内容有很多。

例如，关于见面间隔，我们可以这样解释：距离上次见面时间越久，越容易进来其他工作，因此也就越不容易再见。

收到邮件后一直不回的话，说明邮件中涉及的事本来就很棘手，不能简单地给出回复，之后可能就越来越难回复了。

但是，像这样的个别说明无法证明各不相同的 4 种行动中存在统一的 $1/T$定律。在人类和社会的行动中，或许存在更普遍的机制。

3.7　$1/T$定律与 U 分布相同

实际上，上述 $1/T$定律可以从第 1 章介绍的 U 分布中导出。

人的行动，是每分每秒的无数行动选择不断积累的结果，不管人的个性和处境如何，都遵循统一的统计规律——U 分布。随着行动的不断积累，资源分配会出现不均，行动会完美地遵循数学定律。没想到，人的行动竟和空气中分子的能量分布遵守同一个公式。

我们可以用方格中的小球将 U 分布转化为一种普遍的模型（图 1-3）。用图表来表示 U 分布时，横轴表示方格中的小球数，纵轴表示小球被装进多少个箱子。在这里，1 个方格代表某人的 1 分钟，小球

数代表 1 分钟内胳膊的活动次数，这样一来，普遍的分布规律就出现了。

但是，我们还要注意小球和小球的间隔。去掉图 1-3 U 分布图中的网格后再看一下各小球的间隔。在该模型中，我们可以用小球与小球的间隔表示从某项活动（即胳膊活动、收到邮件、与人见面等）发生到下次发生的时间间隔。

在 U 分布中，将小球与小球的间隔作为横轴重新统计，结果正好遵循了 $1/T$定律。具体来说，将相邻小球的间隔（T）作为横轴，将该间隔的发生概率作为纵轴，则纵轴与横轴成反比（假设横轴数值是原来的 10 倍，则纵轴数值变为 1/10），即 $1/T$。

在第 1 章中，我们统计了方格中的小球数，发现其分布曲线是逐渐下降的（指数函数），并称之为 U 分布。而现在，我们统计了相邻小球的间隔（距离），发现其分布成反比关系（$1/T$），也称之为 U 分布。事实上，虽然着眼点（统计的量）不同，实际状况却是相同的（章末注 1）。

3.8　记述人类行为的方程式

在第 1 章中，为了制作 U 分布，我们首先随机分配了小球，然后随机移动了小球。实际上，如果我们一开始就按照 $1/T$定律来分配小球间隔，就能直接生成 U 分布。

这种直接生成人类活动(胳膊活动、收到邮件、与人见面等)分布规律的公式，就是我们一直在寻找的人类行为的基本方程式[5]。

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t}=-\left(\frac{1}{T}\right)P(t)+F(t) \qquad (3.1)$$

在这个方程式中，P（t）指时间t内，上一个活动（胳膊活动等）发生后，下一个活动没发生的概率（累积概率）。方程式左边是P（t）除以时间，右边是P（t）乘以时间t内下一个活动发生的概率（概率密度），前面再加一个负号。它表示的正是活动的 Generator。T表示自上一个活动之后经过的时间。F（t）是外部施加的力，会导致行动偏离 U 分布。由于方程式右边有 1/T，因此称为 1/T方程式。

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t}=-\left(\frac{1}{\tau}\right)P(t) \qquad (3.2)$$

如果活动（胳膊活动等）是随机发生的，不遵循 U 分布，那么它遵循的就是第 1 章和本章开头所说的泊松分布（近似于正态分布）。泊松分布的方程式早已被导出，如下所示。

在这个方程式中，τ是表示时间的常数，代表活动的平均时间间隔。但是，这个式子与身体运动和见面间隔的实测结果并不一致。

我们来对比一下与实测结果完全一致的 U 分布（3.1）方程式和不一致的泊松分布（3.2）方程式。在泊松分布（3.2）方程式中，右边有一个常数τ。它表示下一个活动都是以一定的频率发生的。换言之，这意味着时间的流逝是固定的。而在（3.1）的 1/T方程式中，时间的流逝随着时间 T 逐渐加快。这是用数学方式来表示 1/T定律——行动持续得越久越停不下来。

重要的是，该方程式成功预测出了经验性定律。也就是说，对于埋头工作时时间就会过得很快(这在心理学中称为“心流体验”，后面会详细说明)，不回复的邮件会愈发难以回复，不见面的人会愈发疏远（“去者日以疏”）等现象，我们可以进行定量预测了。

我们有时会觉得时间的流逝不尽相同，而 1/T方程式为这种主观感觉赋予了科学依据。当今时代，我们通常认为钟表是将时间定量化的唯一手段。但是在古希腊，有两个表示时间的词将时间明确分成了

两类——机械、物理上的时间流逝称为“克罗诺斯时间”，而人类内心、主观感受上的时间流逝称为“凯罗斯时间”。如果说对应钟表时间的克罗诺斯时间是科学、客观的，那么人们可能会认为，主观的凯罗斯时间是非科学的。但是，如果用前后两次身体活动的时间间隔来定义时间的话，就可以为主观的凯罗斯时间提供客观的依据。上述方程式客观表示了古希腊所说的主观时间流逝的概念。

然而，严格来说，$1/T$方程式只在没有外部力量和限制的情况下才会成立。就像第 1 章所说，U 分布表示人类行为的自由度，如果因为一些原因受到了限制，那么人类行为就会偏离 U 分布。

表示 Generator 的方程式（3.1）中有一个外力F（t），在外力的作用下，行动会偏离 $1/T$定律。偏离 $1/T$定律的F（t）的大小因人而异，这就意味着，阻碍（或限制）人类自由行动的外力大小因人而异。

3.9　将主观概念转化为客观数值

从 $1/T$方程式中，我们能得到什么信息呢？实际上，该方程式中隐藏的最重要的信息是：集中精力做事的时候，是人类最自然的状态。$1/T$方程式右边的F（t）表示的是限制自由行动的外力。不受外力干扰，就相当于不受限制的自由状态。这种情况下，行动就像上了发条一样，持续得越久越停不下来。这才是人的自然状态。这就相当于日常生活中所说的“聚精会神”“专心致志”“全神贯注”的状态。

物体运动时，如果没有外力作用，就会保持匀速直线运动。这是牛顿第一运动定律。我们将使物体脱离自然匀速运动状态的作用力称为“外力”，这样一来，就可以从定量角度定义物体承受的外力了。这是牛顿第二定律。

我们将行动不受限制的自然状态称为“集中精力”。可能很多人觉得要想集中精力，就要勤奋、努力，但是该方程式表明，集中精力做手头的事，就是我们的自然姿态。

说到这里，想必有人会质疑，这个方程式真的能表示人类集中精力的状态吗？同样地，集中精力的状态因人而异，不能一概而论。

在测量和定量分析人类行为时，上述疑问是不可回避的基本问题，故在此多加说明。我们用大家早已熟知的温度来进行比较说明。

物体的冷热程度可以用温度计来测量。温度计利用了酒精等物质热胀冷缩的性质，通过受热体积膨胀来测量冷热程度。但是细想一下，其实“热”也是多种多样的。

滔滔最上川，炎日入海流。

幽幽林中寺，闲岩入蝉声。

这是松尾芭蕉在《奥之细道》中吟咏夏天的诗句。这两句体现出来的炎热的性质是截然不同的。第一句表现的是，在最上川注入日本海的地方，立于高台之上观赏鲜红晚霞照耀河川时的热；第二句表现的是幽幽密林中，只看到阳光透过树枝，听到蝉声在耳边回荡时的热。古往今来，日本的歌人、俳人都在表达各种形式的热（温度）。

现实世界中，既有梅雨季节的闷热，也有沙漠的干热。既有加热后金属的热，也有沸腾的水壶中冒出水蒸气的热。热的种类很丰富，即使同为 30 摄氏度，有时是不堪忍受的热，有时却是较为舒适的热。

毫无疑问，现实世界中存在各种各样的热。我们一方面承认这点，另一方面却用温度计将酒精的膨胀量视作温度，用一个单一的尺度将“热”转化成了数值。其实这只不过是现实世界中“热”这一概念的冰山一角。

尽管如此，对温度的定义和测量仍然意义重大。如果没有温度这个共通的尺度，对受温度影响的机器进行设计与调控等工作就将寸步难行。在完全没有尺度的情况下，创造一个客观可测的尺度，哪怕这一尺度不能全面反映人类所体验的丰富现实，也可以在论述各种变化时提供一种共通的语言。而且有了测量尺度之后，那些无法用尺度测量的差异反而更加明确了。

在利用温度计的初期实验中,当时的科学家们并不知道温度计测量的究竟是什么(托马斯·库恩《必要的张力》)，但可以肯定的是温度计与冷热程度相关。但是，温度计测量的温度与人们感知的“本来的准确温度”有着明显的不同。温度计的测量值相同，人们感受到的温度却完全不同的情况时有发生。因此，我们不禁会觉得，温度计捕捉的不是准确的温度，而是某种复杂难懂的事物。

但现在我们知道了，真正复杂难懂的是我们的感觉。温度计只不过朴素地表达了构成物质的微粒——原子运动的激烈程度。

为了接受温度这一概念，我们需要扭转看法。也就是说，我们必须认识到，并非感觉易懂而温度计复杂难懂，恰恰相反——温度计是客观易懂的测量尺度，而感觉是模棱两可的复杂现象。在 300 多年前的 17 世纪，关于温度，人们的看法发生了逆转。

就像人类克服温度方面的思想障碍一样，要想将“集中精力”定量化，也必须克服思想障碍。更何况在集中精力方面，基于个人经验的主观感觉已经遍布各个领域，深深地在人们心里扎了根。不管是业务中、生活中，还是法庭上，人们都在利用经验感觉行事。就像对温度的感觉与温度计的测量值并不吻合一样，对集中精力的科学测量值与经验感觉也不完全一致。

但是，测量值比感觉更客观，有着更加可靠的依据和更为坚实的科学基础。测量值与感觉不同，它可以被大量记录并提供参考，还可以从数学角度来表示数据的变化和规律性。

如果我们确立了新的测量值，就可以通过微分定义变化，通过积分定义积累，还可以将几何学的结构定量化。这意味着，科学在过去数百年来打造的各种工具终于可以派上用场，进步的速度与过去相比将不可同日而语。而且最重要的一点是，一旦对该测量值达成共识，人与人之间就可以共享此概念，进而根据共通的概念开展有效对话。

温度计的历史告诉我们，不应追求测量值与感觉保持一致，而应发现有科学基础的测量值，并确立其理论依据。然后，以此为基础创造一个新的世界观。

通过 $1/T$方程式，我们发现了人类最自然的状态——集中精力的重要性。难道迄今为止没有人发现它的重要性吗?

当然不是。集中精力对人类来说是一种特别重要的状态，这在心理学上早已有所认识，而且也一直在进行定性研究。那么，我们在此构建的定量方程式是如何进一步深化前人的智慧的呢？下面我将进行说明。

3.10　测量最优体验=心流

匈牙利心理学家米哈里·契克森米哈赖教授将人集中精力、全神贯注做事的状态称为最优体验或心流状态[6]。换言之，就是能感受到自己正在做的事情的价值，能发挥自己的能力并享受其中的体验或状况。这里说的最优体验，和我们得出的“人最自然的状态”的结论有异曲同工之妙。

心理学家研究了多年人内心的主观体验。他们利用的方法论是通过问卷调查（即让被测人用数字来回答问题），将人们的感受转化成数字。这样一来，就通过主观的回答将调查对象“心理”转化成了定量数值，并逐渐形成了一门学问。而由新的可穿戴式传感器通过身体运动测出的大数据，客观而定量地揭示了人类行为的特点。

两者就像是从不同的方向攀登同一座山，当两者的智慧在山顶会合，就将客观且科学的知识与本人的主观真实感受融合在了一起，从而使我们能够理解人类的行动。

契克森米哈赖教授曾研究过一些在工作、运动和个人兴趣等方面表现杰出的人，这些人都不约而同地说出了相同的体验。他们都说，自己可以专心致志地做手头的事，忘记时间的流逝，而且还会意识不到自己的存在，感觉自己和周围融为了一体，可以随心所欲地控制对象。

处于最优体验即心流状态时，人就会体会到一种愉悦感和充实感。另一方面，我们发现，当所注意的对象瞬息万变，令我们无法集中精力时，就浪费了我们的精神能量，使我们难以体会到愉悦感和充实感。

心流体验的频率因人而异，有的人经常在生活和职场中体验心流状态（占 1 天的 40% 以上），也有人几乎没有体验过心流状态（还不到 1 天的 10%）。

我们认为心流体验的频率与 $1/T$方程式之间可能存在某种关系，并为了验证其对应关系而进行了实验。该实验是和心流概念的倡导者，契克森米哈赖教授共同实施的[7]。

契克森米哈赖教授为了测量心流状态，想出了一个特别的方法，并一直将其运用于实验中。该方法叫做“经验提取法”，现已广泛应用于心理学实验。

在经验提取法中，被测人随身携带传感器和手机。实验者以平均约 90 分钟 1 次的频率，让传感器（或手机）随机响起哔声，然后让被测人就哔声响起瞬间的自我状态回答一份简单的问卷。这种方法的关键在于，在出人意料的时候突然“扣下扳机”，留下对当下体验的新鲜记录。所谓问卷的问题，具体指的是：

刚才你做的事困难吗?

刚才你发挥出自己的能力了吗?

让被测人用 1 到 5 的数字（5 非常符合，1 完全不符）回答这些问题。第一个问题将手头工作的“挑战度”转化为数值，第二个问题将“能力发挥度”转化为数值。组合这两个问题的数字，就可以将被测人的状态划分为 2×2=4 种（图 3-2）。

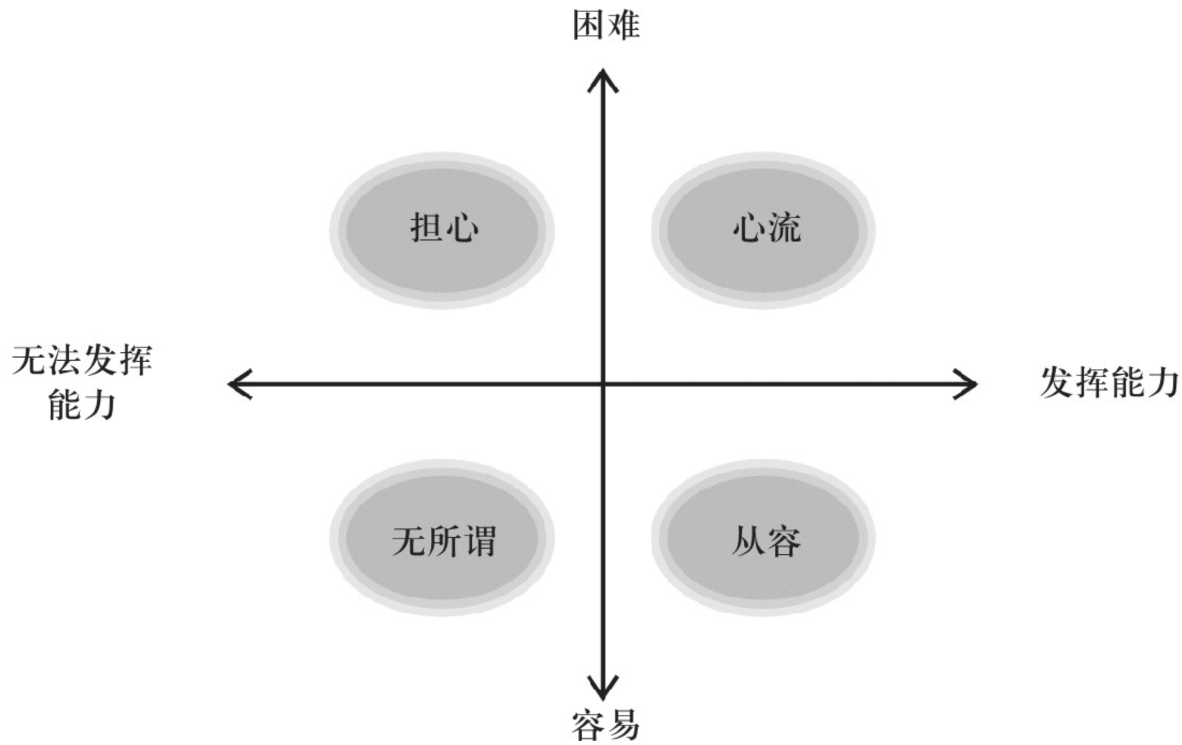


图 3-2 心流、从容、无所谓、担心这 4 类与能力发挥度、挑战度的关系。

挑战难度大，能力难以发挥的状态为“担心”；挑战难度小，能力得以发挥的状态称为“从容”；平衡“担心”和“从容”这两种状态，集中精力做手头工作的状态称为“心流”；能力发挥度和挑战度都低的状态称为“无所谓”。在这里，我们用被测人在特定期间内回答的平均值来划分挑战度和能力发挥度的高低。这样一来，我们就可以将数值偏高的人与数值偏低的人的差异标准化。

通过为期两周的测量，可以获得哔声响起时（几十次）的数值和当时被测人属于担心、从容、心流、无所谓中的哪一状态。

为了调查这种心理体验数值和身体运动的对应关系，我们为被测人戴上了可穿戴式姓名牌传感器，调查他们的身体运动、是否与人见面等情况，同时通过经验提取法展开问卷调查。通过这 2 周的传感器数据，我们获取了被测人身体运动的特点和与人接触的频率等信息。我们利用这些数据，制作了表示各种身体运动特点的指标，并调查了该指标与此人心流状态的相关性。

我们特别关注了身体运动的持续性。$1/T$方程式告诉我们，人最自然的状态，是身体运动持续越久越停不下来的发条状态。遵守 $1/T$定律，也就意味着此人的身体运动遵循 U 分布。而且在第 1 章中说过，身体运动遵循 U 分布时是熵最大化的状态，也可以说是一种行动自由不受限制的状态。如果将这种自由行动的状态与心理学的最佳状态结合起来，说不定可以验证内心和身体的密切关系。

而实验结果也证实了这一点。快速身体运动（2～3Hz，即接近 240～360 次/分的步行节奏）的持续与最优体验（=心流状态）的频率密切相关。

具体地说，我们比较了前后两个 5 分钟的快速身体运动的频率，发现心流频率高的人，身体运动的频率变化较少。容易进入心流状态的人，更倾向于持续进行快速的身体运动。这表明，持续快速的身体运动会促使人集中精力做手头的事，而集中精力的人的身体会持续快速地运动。

顺便一提，经过确认，我们还知道心流体验时不一定会接连产生身体的运动。心流体验的过程中，也并不需要持续或快速的身体活动。

内心的活动和身体的活动就这样联系在了一起。实际上，这个发现为我们掌控自己的人生提供了重要的启示。

人最想又最难掌控的是自己的内心，尤为重要的是是否能够享受其中。我们发现，心流状态表达的是是否享受其中的主观感受,而这种体验频繁与否，与可用传感器测量的身体活动密切相关。在工作和生活中体会到愉悦感和充实感的人，身体运动的持续性也比较高。大家都渴望的愉悦感和充实感，本来是一种抽象无形的感受，但现在成了具体可见的东西。

通过创造一个促使身体持续快速活动的环境，或许就可以在工作和生活中获得愉悦感和充实感。并且，如 1/*T*方程式所示，这种身体活动不是特殊的，而是人最自然的状态。

然而，由于工作、社会方面的限制和责任，我们可能会将这种自然的状态抛之脑后。但是，根据我们传感器的测量结果，就可以客观地了解自己的状态。具体来说，根据传感器每天上传的数据，我们可以确认遵守 1/*T*方程式的持续活动的频率和心流状态的频率。每天确认体重计的测量，会影响我们的饮食生活。同样地，对身体运动的测量则有望成为使我们的人生保持在高度自由与自然状态的新技术。笔者在看到自己的传感器测量结果后，会有意识地增加 2Hz 以上的快速运动。我们还证实，这一结果与自己对每天生活的满意度有关。快速运动增加后，心流状态出现的频率也明显增加。为此，我有意识地做了一些具体的努力，比如对话时，能站着就尽量不坐着。因为这样的话身体易于活动，容易进入心流状态。在工作停滞时，我会有意识地在办公室里走来走去，以增加超过 2Hz 的身体运动。

自此，我们开拓了一条新的道路——通过控制身体，来控制内心。

注 1

U 分布的特点是，从对象事件的频率（或密度）来看，遵循指数函数（数学上称作指数分布）；另一方面，从对象事件的间隔来看，遵循“1/*T*定律”，事件间隔的发生频率遵循“幂分布”（即让事件的发生

遵循 $1/T$定律，统计事件间隔的发生频率，结果呈幂分布）。前面说的巴拉巴西教授和中村先生之所以在论文中称其为幂分布，就是这个原因。

但是，一种是事件间隔呈幂分布，另一种是对象事件的频率与密度呈幂分布，这两种现象的意义截然不同。当事件间隔遵循幂分布时，我们仔细看一下事件密度就会发现它遵循的是指数分布。笔者将这种实际状况称作“U 分布”。

在物质世界，大家都知道平衡状态下的物质遵循“玻尔兹曼分布”（能量呈现指数分布）。但是，在遵循玻尔兹曼分布的实际物理状态（例如粒子的分布）下，如果统计一下各种物理量（例如与旁边粒子的距离）的分布情况，那么除了指数分布以外，还会出现各种各样的函数分布形式。但是，遵循玻尔兹曼分布的实际物理状态这个前提是不变的。从这个意义上来说，玻尔兹曼分布不是数学上的指数分布的别称，而表示了一种实际物理状态——平衡状态。因此，统计分布相关文献中的讨论便容易陷入混乱。一种“分布”是指数学上的函数形式，另一种“分布”是指物理上的实际状态，人们没有认识到这两种情况的区别（这个解释稍显专业）。U 分布将玻尔兹曼分布普遍化，应用于人类等非物质现象中，它指的不是数学上的函数形式，而是物理上的实际状态。可能有人误以为 U 分布就是指数分布，为了避免这种误解，特作说明。

第 4 章
认真面对运气

4.1　偶然是不可控的吗

我们在第 2 章中介绍过，业务生产力表面看来是由个人能力和性格决定的，但实际上深受周围人的身体活跃度的影响。

这时，“活力”这个乍一看有些陈旧的词语进入了我们的视线。近 30 年来，在崇尚公司运营的理性与逻辑的社会潮流之中，“活力”更是成为无人问津的词语之一。在这里，我们通过解析科学数据，重新认识了活力的意义。

本章中，笔者再举一个与“活力”有着相同遭遇的词——“运气”。

涩泽荣一奠定了日本近代经济的基础，他在其著作《论语与算盘》（1916 年）的“成败与命运”一章中，认真论述了怎样看待运气、如何受到好运加持。

再举一个例子。大文豪幸田露伴写过一本《努力论》（1912 年）。他从序章开始，就以“命运和人力”为题，论述了怎样看待努力与偶然交织的人生、如何生活等。

曾经，运气被如此认真地讨论过。然而，不知从何时起，人们不再认真面对运气了。人们越来越多地将“运气”一词束之高阁，不予理

睬。但是，古往今来，我们的人生和工作能否成功，都受到运气的巨大影响。

那么，在现代的商业学校和经营者培训中，人们是否认真讨论过运气呢？实际上，人们不仅不会讨论，甚至认为“运气”与迷信和占卜挂钩，是应当在理性的经营中被排除在外的概念。

在这里，我们将“运气”定义为在人生和社会中根据概率发生的好事，即在人生和商业领域中，人们期望发生的一种概率现象。

这样一来，我们可以将运气看作科学家们多年苦心研究的一种统计现象。统计学原理在方方面面都发挥着作用。科学研究表明，在既没有设计图，也没有指示的情况下，水达到 100 摄氏度时会沸腾，碱基会相互聚集，形成美丽的双重螺旋结构，这些现象都是随机过程的结果。那么，这些科学理论能否用来认识人生和企业的命运，并将其引向更好的方向呢?

既然遇见好运是一个概率性事件，那我们可能无法预测或猜中运气出现的时机。

但是，当概率现象反复发生时，我们就可以预测并控制现象的发生频率。也就是说，如果我们将遇见好运视作多次反复发生的现象，而非一次性现象，就可以对其进行预测了。例如，掷 1 次骰子能否出现偶数，是由 1/2 的概率决定的，那么当我们掷了 1,000 次骰子，想知道偶数出现了多少次时，就可以预测大约是 500 次。如果掷骰子的次数不断增多，其误差就会越来越小。随着试验次数N的增加，运气是否也可以成为能够科学预测的现象呢?

在我们的人生中，每天都在不断地遇见好运，遇见好运的机会每天都在重复。当我们选定一个期间，在此期间内，随着试验机会N的增加，对遇见好运的频率的预测精确度也会提高，进而有可能控制运气。结论就是，科学的方法，尤其是统计物理学的方法论是一个强有力的武器，有助于提高我们遇见好运的概率。

人生和社会中发生的事几乎都是必然与偶然交织的产物，抛却偶然因素的事件几乎为零。因此，很多事都是概率性事件，这时运气就显得尤为重要。

但是，我们往往倾向于用“偶然”和“必然”这两个对立概念将事物一分为二，即简单地分为不可控的偶然现象和可控的必然现象这两类。然后我们就致力于控制必然现象，轻易地放弃了不可控的偶然现象，并且还以为这是理性的判断。

但实际上，对于伴随偶然因素的现象，其概率也是可以控制的。打棒球时，击球手不一定每次都会击中。但是，我们通过起用打击率为 30% 的击球手，而非 20% 的击球手，就可以提高上垒的概率。不过只看一次打席的话，也会发生 20% 的击球手击出安打，而 30% 的击球手没有打到的情况。但是，随着打席数的增多，安打数就会出现明显的差异。对于这种包含偶然因素的现象，如果我们从一开始就放弃了提高概率的可能性，也就等于坐失了诸多良机。

4.2　运气源于与人的相遇

前面我们将运气定义为凭概率发生的好事，如果具体到商业领域重新定义一下的话，可以认为是凭概率遇见拥有自己所需的知识、信息和能力的人。

苹果公司的创始人史蒂夫·乔布斯戏剧性地改变了计算机和人类之间的关系，所以会名垂青史。但在这里，“运气”也发挥了很大的作用（以下内容参考了沃尔特·艾萨克森所著的《史蒂夫·乔布斯传》）。

乔布斯饱尝了失败之苦。为了提高他所创立的苹果公司的经营水平，乔布斯亲自将约翰·斯卡利招进公司。但是由于两人经营方针不一致，乔布斯被迫离开了苹果公司。

就在失意期间，乔布斯参加了一次午餐会，旁边坐的碰巧是诺贝尔奖获得者保罗·伯格，两人谈论起基因重组技术。伯格说，生物学中的实验非常艰辛，有时甚至要花费数周的时间。乔布斯问道：“用计算机进行模拟实验怎么样呢？”伯格教授解释说，能做这种模拟实验的计算机太过昂贵，大学是买不起的。就是在那个瞬间，乔布斯察觉到了一种很大的可能性，顿时两眼放光。

在那之后，乔布斯创立了 NeXT 公司，专门生产面向大学的工作站电脑。NeXT 公司本身在商业方面算不上大获成功。但是，在将近 10 年后，苹果公司购买了 NeXT 公司的优秀软件，随后收购了 NeXT 公司，乔布斯以此为契机，重新返回了苹果公司的经营层。该事件与后来乔布斯和苹果公司突飞猛进的发展息息相关。

总结一下，乔布斯在午餐会上与伯格的相遇以及由此引发的伯格的言论，推动了他后来的飞跃发展，即相遇招来了好运。

日本汉字学家白川静指出，自古以来，人们认为幸福四处巡游，因此用了表示“搬运”的“运”字。事实上，很多时候，运气都是通过与人的相遇得到的，包括乔布斯的例子在内，古往今来的传记和体验记录中，都有很多相关记载。

4.3　将运气和相遇转化成理论和模型

如上所述，人与人之间的相遇和对话，有助于我们遇见好运。但是，谁在何时会给自己带来好运，是无法提前设计的。正是因为无法提前设计，才称为运气。

前面已经介绍了记录人与人之间见面情况的可穿戴式传感器技术。姓名牌传感器中装有红外线收发器，将传感器戴在脖子上，就可以随时记录所见之人和相应时间。将这些记录统计起来，就可以制作出“社交图谱”（关系图谱），表示谁和谁有关系、谁和谁在对话。社

交图谱用图（由点连成线）来表示人与人之间的关系，当我们想在SNS 等社交平台上俯瞰朋友关系网时，就可以使用社交图谱。过去，我们很难制作出表示现实世界中的人际关系的社交图谱，但是借助可穿戴式传感器，我们就可以记录人与人之间的见面情况，然后将这些数据输入计算机并制作出社交图谱（卷首插图 4）。

当我们转变一下看法，借助这一传感器就可以将遇见好运的机会定量化了。接下来说明一下定量的方法。

刚才举的史蒂夫·乔布斯的例子中，乔布斯与伯格教授的会面堪称一次“幸运邂逅”，它带来了后来的巨大发展。通过与人的相遇，也许我们可以获得启发，发现对问题的另一种看法。对方可能会介绍我们需要的书籍，甚至介绍我们认识其他人。传感器捕捉的就是这样的机会。

当然，与人见面并不意味着一定会发生什么好事，两者之间没有必然的联系。但是，与人相遇会提高我们抓住好运的概率。借助该传感器，我们可以用数学将这种机会定量化。

实际上，我们并不知道运气存在于何处。自身的需求和周围的供给时刻都在变化。即使供需偶然重合，也只有在两者的共通点成为话题时，运气才会奏效。

下面让我们试着将这种状况转化为理论。假设对自己有用的信息和有能力的人按照一定的概率，随机分布在自己的周围。

在这种情况下，每当自己与人见面时，都有遇见有效信息和能力的可能（或概率）。简单来说，就是你见面交流的人越多，遇见好运的概率就越大。

但是，在见面的人中，既有人脉广、消息灵通的人，也有很少与人对话、信息渠道有限的人。我们可以预见，前者为你带来有效信息（你的运气）的可能性相对较高，而后者的可能性较低。

要想预测见面的效果，可以估量一下对方的人脉情况，即累计一下这些人的见面人数。这与如果连同你朋友的朋友都算在内，你的人

际圈将涵盖多少人的调查是一个道理，我们称之为 2 步以内的“到达度”（图 4-1）。2 步以内的到达度是衡量是否容易遇见对自己有用的信息和能力（遇见好运）的指标。

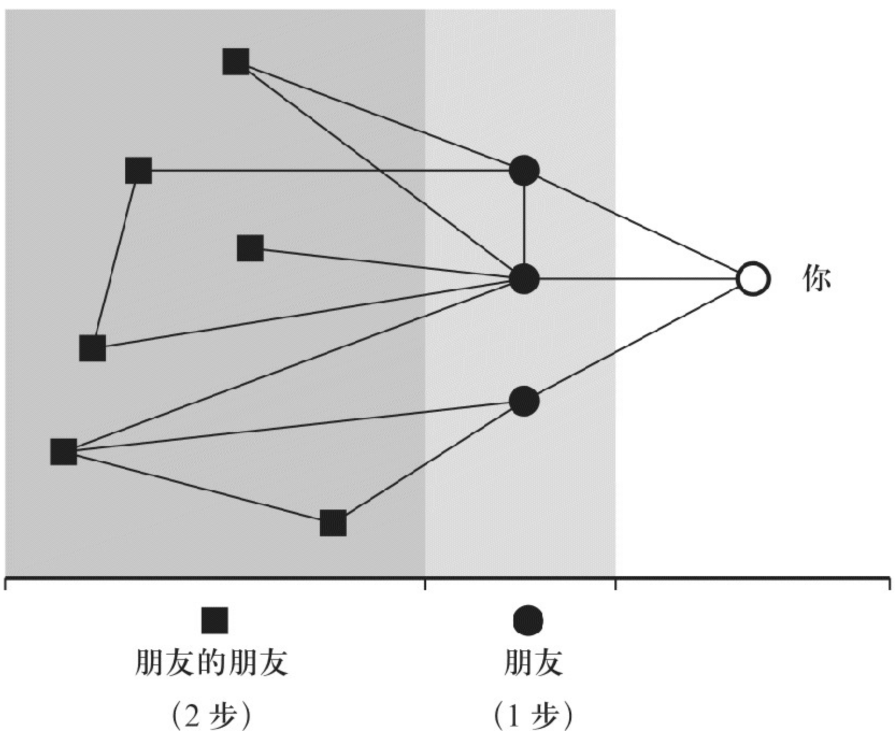


图 4-1 表示 2 步以内的到达度的图。该图中，你的朋友和朋友的朋友一共有 9 人，因此 2 步以内的到达度为 9。

假设你有 3 个朋友，这 3 个朋友分别有 2 个新的朋友，那么你的到达度就是 9 人（直接朋友 3 人加上间接朋友 2×3=6 人。当然你的 3 个朋友也可能互为朋友，但由于没有产生新朋友，所以这种情况不算在内）。因为连同朋友的朋友在内，你和 9 个人产生了联系。如果你同样有 3 个朋友，但是他们分别认识 10 个人的话，那么你的到达度就是 33 人。前后两种情况中，你的直接朋友的人数都是 3 人，但是后一种情况中，你通过他人获取信息与能力的可能性要大得多，这种差异就体现为到达度的差异。

当然，像朋友的朋友的朋友这种 3 步以上的关系，也可能为你带来有用的信息和能力（运气）。但是比起 2 步的关系，信息搬运的难度提高了，因而带来有用信息和能力的概率也就降低了。

因此，2 步的到达度（2 步即可到达的人数）可以作为衡量运气好坏的指标（后面说的到达度，就是指 2 步以内的到达度），表示你遇见好运的概率。而且，我们可以借助传感器将这一概念定量化。

4.4　到达度真的是衡量运气好坏的指标吗

前面说的归根结底只是在假设成立的情况下的理论性考察。用这种方法定义的指标，表示的真是运气的好坏吗?

我们已经得到了证据，可以证明这一点。我们和麻省理工学院（MIT）共同调查了某企业的部门针对顾客有关 IT 系统的咨询，为顾客提供报价方案的经营活动[1]。面对顾客的咨询，约 30 名员工在约 1 个月内提出了 900 多个报价方案。我们从运气的角度上调查了制定报价的工作是否顺利。我们为员工戴上了可穿戴式传感器，通过他们的交流情况来判断运气的好坏。

与顾客的交易中，既有可以机械报价的简单咨询，也有无法轻易答复的复杂要求。其中，面对复杂要求时,我们必须借助周围人的信息和能力。如果顾客的要求超乎预料，不知道本公司的产品能否满足需要，我们就无法将报价的回复方法写进报价手册了。在这种情况下，我们无从得知回复报价所需的信息和知识在哪里。因此，左右结果的便是遇见所需信息和知识的运气了。

经调查发现，在这项工作中进展顺利的人（从接受复杂的报价要求到答复报价的平均时间短的人）有一个共同特征。如果单纯是交流的朋友比较多，那么他的工作不一定能顺利推进。也就是说，这样的人有可能只是人脉广，而不能有效利用周围潜在的信息和能力。

实际上，在该工作中进展顺利的人的共同特征是到达度高。前面说过，到达度指的是包括朋友的朋友（2 步）在内,可以与自己产生联系的人数。我们在前面讨论过，包括朋友的朋友在内，我们可以将获取自己欠缺的信息和能力的这种能力转化成定量数值，作为衡量运气好坏的指标。对于顾客提出的意料之外的咨询，到达度高的人虽然也不知道答案，但是他们得到答案和启示的概率较高。如果这样的事就发生在自己身边，那看起来就是此人的运气好。运气好的人可以处理始料不及的复杂问题，并顺利开展工作。

我们观察每件事情的时候，会发现运气好的人有时也会不顺，运气不好的人有时也会一帆风顺。但是，如果看一下多件事情的统计情况，就会发现运气的好坏，大大左右着人们对意料之外的复杂事件的处理。反过来说，工作进展不顺的人，他的到达度就低。也就是说，由于运气差，所以处理不好那些手册里没写的预料之外的情况。

4.5　运气好的人在组织中处于什么位置

借助可穿戴式传感器，我们不仅可以将到达度转化成定量数值，还可以直观地查看运气的好坏。这是因为，通过可穿戴式传感器中嵌入的红外线传感器，可以将你和别人见面的数据记录下来，如前文所述，我们可以根据这些数据绘制出社交图谱。

例如，用白圆圈（○）表示你自己，用黑圆圈（●=1 步的见面者）表示你遇到的人,用方块（■=2 步的见面者）表示你遇到的人所遇到的人,见面的人们之间用线连接。这样一来，你的白圆圈附近就分布有黑圆圈，再往后还分布有方块（图 4-1）。

就算是见面了，如果只是说了几句话，或者打了个招呼，那双方拥有的信息和能力很少会帮到彼此。因此，只有当两人的见面时间超

过一定的基准值（为引出有用的信息和对方的能力，我们设定了一个经验基准值，即每周要有 15 分钟以上的对话）时，才能用线连接。

运气好的人的到达度高，那么在社交图谱中会呈现出什么特征呢？从社交图谱来看，到达度（2 步以内即可相连的人数）高的人周围有很多人。相反地，到达度低的人周围的人较少。

通过这种人与人之间的关联，你从周围人拥有的信息和能力中受益的可能性就会提高。根据简单的算法，就可以用电脑描绘社交图谱——到达度高的人位于中央，到达度低的人分布在周围。这样一来，容易从周围人的信息和能力中受益的人（到达度高、运气好的人）就位于中央。反过来，鲜少从周围人的信息和能力中受益的人（到达度低、运气差的人）就分布在四周，且周围的人很稀疏。

迄今为止，我以 100 多个实际存在的组织为对象，观察了这一社交图谱。我从中发现，位于社交图谱中心的人，不一定就是职位高的人。

我们来看一下卷首插图 4 中软件开发组织的社交图谱。高桥部长（化名）位于左侧，偏离中心。由于高桥部长是最近外聘的新部长，与组织的关联还比较少，只有若松科长等 3 人与高桥部长相连。因此，高桥部长被挤到了周边，到达度比较低——只有 5 人，也就是说从周围的能力和信息中受益的概率较低（运气不太好）。要改善高桥部长的运气，就必须让他和组织中真正的关键人物相关联。这样的话，他应该会逐渐移向组织的中心。

所谓真正的关键人物，指的是组织中的哪个人呢？这在图中一目了然，即喜多先生。喜多先生位于图的中心，从领导若松部长到下属及同事的广泛范围，与很多人都相关联。从数值来看，喜多先生的到达度为 19 人，是高桥部长的 3 倍。

因此，为了改善高桥部长的运气，增加他与喜多先生的对话是一种极为有效的方式。我们假设高桥部长非常忙碌，难以快速增加直接跟自己对话的人。但是，高桥部长仅通过和喜多先生对话，就可以大

幅提高到达度，即运气。从数值来看，高桥部长现在的到达度是 5 人，仅通过增加与一个人的对话，到达度就能提高至 13 人。将衡量运气好坏的指标，即到达度提高至 2 倍以上，从组织成员的信息和能力中受益的可能性就会大大提高。

这样一来，我们就可以依据定量测量结果，在各自的时间限制下，有效地提高自己的运气=到达度。依据传感器的测量结果，系统可以提示我们每个人与谁进行对话最有效，从而提高我们的到达度。

在接下来的论述中，笔者将到达度与运气的好坏视作相同的概念，将带引号的“运气”作为到达度的同义词。

4.6 领导的指挥能力与现场的自律并不矛盾

很明显，组织的兴衰深受“领导的运气”的影响。

我们用到达度来评价各种各样的组织中“领导的运气”，并对如何创造更好的组织进行了研究。

这里我们要注意一点，不是领导的对话者越多越好。领导的对话者增多，虽然理论上说到达度的数值提高了，但从现实角度出发，考虑到领导的时间限制，对话者的增加终究是有极限的。

实际上，有一种方法可以提高领导的到达度，且完全不用增加领导的对话者和对话时间。

为了找到该方法，我们对比了各种组织的数据，观察了运气好的领导（到达度高的领导）身上有没有什么共同特征。当然，每个领导的到达度的高低也大不相同。我们调查了领导的到达度高的组织，观察他们的社交网络有何特点。表面看来，与领导有直接关联的人数，直接关系到领导的到达度。但实际上，两者之间几乎没有相关性。

重要的是组织成员之间的关联。我们调查了领导的到达度与各种指标的相关性，发现成员间的关联呈三角形的情况较多时，该组织中“领导的运气”就好。

这里的三角形指的是将朋友与朋友相连后呈现出三角形。例如，选取你的两个朋友（A 和 B），如果这两个人也相互认识的话，那么你和 A、B 就构成了三角形。反过来，虽然你认识 A 和 B，但是 A 和 B 相互并不认识，那么你们就无法构成三角形（图 4-2）。

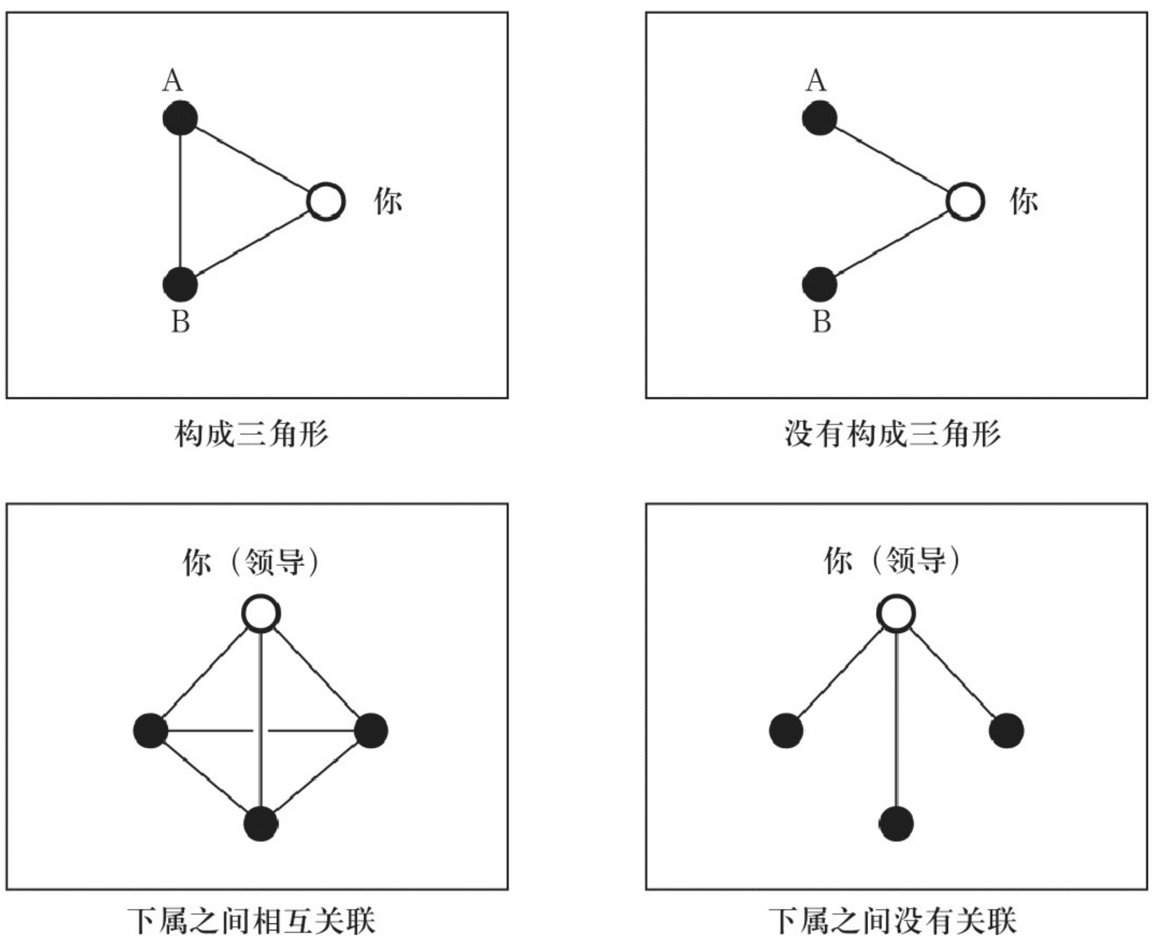


图 4-2 上图表示社交图谱中三角形的形成。如果你认识的人之间也相互认识，就会在你的周围形成三角形。

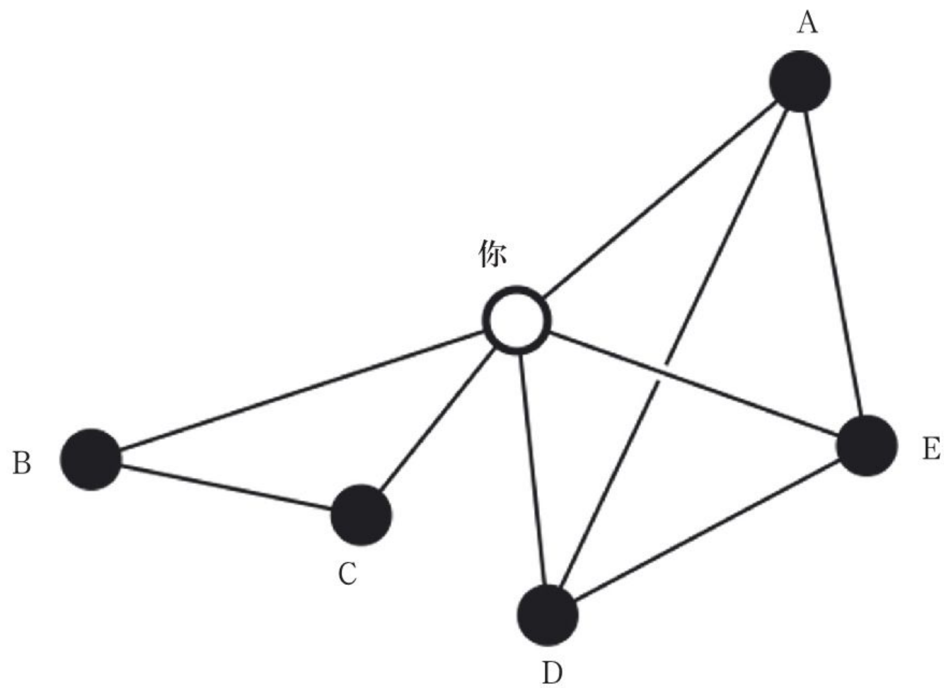
下面的情况在组织中十分常见。假设你有 5 个下属，你和这 5 个下属相关联，他们以你为中心呈放射状分布。但是，仅仅如此是无法构成三角形的。要想构成三角形，5 个下属之间必须相互关联，也就是说下属要直接和其他下属对话。除了领导和下属的关联以外，还要有其他的关联才能构成三角形。

当成员间的关联呈现三角形时，会有什么变化呢？首先，我们考虑一下完全没有三角形的情况：你和 5 个下属的对话呈放射状，下属之间没有关联。

这时，假设下属 A 的工作出现了问题，而其实这个问题可以用成员 B 拥有的信息来解决。当你在办公室时，A 找你商量怎么解决问题，你可以从 B 那里问出所需信息，然后告诉 A。但是，如果你出差了，那么等你回来后 A 的问题才能解决。因为 A 不知道 B 拥有的信息，所以无从询问。

如果 A 和 B 平时直接交谈过呢？也就是说形成了“你—A—B”的三角形会怎样呢？这样的话，即使作为领导的你不在，A 也可以直接问 B，问题就解决了。

在组织的社交网络中，如果三角形多的话，那么即使领导不直接介入，在现场主动解决问题的可能性也会提升，即现场的运气会得以改善。三角形多也就意味着现场的成员之间是相互关联的，即可以用三角形的数量来定量地表示现场凝聚力的强弱。因此，数一下社交图谱中各成员周围有几个三角形，再取全员的平均值，即可得出凝聚度（图 4-3）。



你周围的三角形数量 =4

A 周围的三角形数量 =3

B 周围的三角形数量 =1

C 周围的三角形数量 =1

D 周围的三角形数量 =3

E 周围的三角形数量 =3

凝聚度（各成员周围的三角形的平均数量）为 15÷6= 2.5

图 4-3 上图表示的是组织“凝聚度”的计算方法。数出各成员的周围有几个三角形，然后取平均值，即得出凝聚度。

我们整理了一下各种组织的数据，发现在到达度高的领导带领的组织中，衡量现场能力的凝聚度指标也高。再揣摩一下，就会发现这是件出乎意料的事。

由于到达度高的领导在 2 步以内就和很多下属相关联，因此他不仅可以掌握组织的实际情况，还可以很容易地将指示传达给下属。简言之，该组织的“领导力”强。

但是，有很多人担心，在这种领导力强的组织中，现场能力会变弱。前面所说的凝聚度正是衡量现场能力的指标。因此，有很多人认为“领导力”即“领导的到达度”与“现场能力”即“凝聚度”是不能两全的。然而，我们通过分析大量数据，得出了领导力强的组织的现场能力也强的结论。

按下面这种方式思考的话，就可以理解上述结论了。在社交网络中，三角形表示的是“shortcut”（捷径）。举例来说，假设 A 通过 B 与 C 相连。这时，如果 A 和 C 能够直接关联的话，ABC 就构成了三角形。由此一来，A 与 C 之间就形成了 shortcut。也就是说，原本从 A 到 C 要走 2 步，在形成三角形后，走一步就可以直接从 A 到 C。

当成员间存在很多三角形时，就会产生很多 shortcut。在这样的组织中，领导可以利用 shortcut 在 2 步以内与众多成员关联起来。这样一来，领导的到达度提高，更易利用成员的能力和信息（运气变好）。

我们来分析一下卷首插图 4 的软件开发组织中的这一情况。该组织中，高桥部长的到达度只有 5 人。虽然成员人数少，但是从领导高桥部长到最末端的下属，却有 5 步之遥。

之所以产生这种情况，不单单因为高桥部长只和 3 个人相连接，还因为凝聚度（社交图谱中每个人的周围平均有几个三角形）只有 2.2。平均来说，组织的凝聚度在 4 左右。

我们来看一下喜多先生。喜多先生的关联者多达 9 人，但这 9 人之间并无关联。不只是喜多先生，整个组织的三角形数量都很少。

因此，即使高桥部长不改变自己的关联者，只要在组织中增加三角形，就可以提升自己的“运气”。也就是说，通过提高现场的凝聚力，促进成员间的相互关联，领导就可以有效发挥组织的能力了。

成员间三角形增多的另一个好处是，各成员的到达度会提高。这样一来，成员有效利用周围人的能力和信息的可能性也随之提高（成员的运气也会变好），进而成员的工作也会顺利推进。

我们前面介绍过复杂的IT系统报价业务，在该业务中也可以看到这种三角形，即凝聚度的作用。平均来看，自己周围的三角形多的职员可以用较短的时间回答顾客的复杂要求（召来好运）。

4.7　通过数值化，从语言的束缚中解放出来

如果领导的“运气”变好，成员的“运气”也会变好,反之亦然。领导和现场成员相互依存。

但是，为什么那么多人觉得领导力强的话，现场的能力就会低呢？这是因为人有一个思考习惯，一旦使用“领导力”和“现场力”这种对比鲜明的词，我们就会条件反射地认为这是一个二选一的问题——“领导或者现场”，或“自上而下或者自下而上”。

诸如此类，被二选一的词语束缚的思维习惯随处可见。在关于是“政治主导”还是“官僚主导”的讨论中，人们从一开始就将政治和官僚相互依存的可能性抛诸脑后了。此外，在关于企业的目的是“营利”还是“非营利（公益）”的讨论中，人们也将追求利润和实现公益相互依存的可能性抛诸脑后了。在关于是“工作”还是“个人生活”的讨论中，人们也将工作丰富个人生活，个人生活提升工作成果的共同作用抛诸脑后了。“工作与生活相平衡”本身就包含工作和生活的较量，因而象征着人们的思维习惯被这个词语所束缚。

有关人类的科学定量数据启发我们，除了二选一以外还有别的选择，即实现两者的整合与协调。定量的测量数据超越了人类的认知极限，展示了现实真正的样貌。

4.8　通过控制“到达度”，成功实现组织整合，防止开发延迟

最近，我们经常在报纸和新闻中看到经营整合和企业收购的报道。这是因为，当今时代的全球化商业竞争愈发激烈，只有在业界排名靠前的几家公司才能在竞争中存活下来。

除了企业间的大型合并之外，还有企业集团内部的组织重组，这样一来组织整合就更加频繁了。能否顺利推进组织整合是影响到日本及世界经济发展方向的重大问题。

企业整合对组织及其职员的命运都会产生重大影响。在新公司中，结识新朋友的机会也会增多。以前他们是竞争企业的职员，现在成了你的合作者，而且你还可能与他的顾客和伙伴产生联系。从这个意义上来说，整合是一个机会，我们应该积极利用。

另一方面，在环境、组织和业务急剧变化之时，有的人与他人的关联较少，找不到新的机会。对内调整也可能会耗费巨大的能量，伴随着风险。如果职员无法有效利用自己周围发生的变化（反而降低了运气），那么企业整合很可能面临失败。

将新的组织图画在纸上很简单，但是要把活生生的人类群体整合起来并不简单。要想将背景和文化不同的群体整合起来，不管是职员还是经营者都要竭尽心力。而且不是所有事情都可以提前设计好，有很多事需要做了之后才知道。因此在很多情况下，结果都不尽如人意。

在前面的论述中，我们一直在强调周围人的能力和信息，而当结果不尽如人意时，控制利用能力和信息的概率（即“运气”）的技术就会发挥作用。接下来，我们将通过有关企业内部组织整合的具体实例，说明对“运气”的控制发挥了怎样的作用。

在该公司中，曾经有两个不同的部门分别开发和制造两个产品系列。这两种产品在技术上存在很多共通的地方，因此公司设想，如果将这两个组织合并起来，更有效地利用开发资源的话，理论上可以增强竞争力。但是，这两个部门存在历史和文化差异，想要合并并不容

易。公司的高层担心这一点，于是委托我们利用可穿戴式（姓名牌型）传感器进行观测与支持。

首先，在合并之前，我们让成员持续佩戴了可穿戴式传感器。然后我们观察了两个组织刚刚合并后的社交图谱（卷首插图 5），可以看到代表不同组织的两种颜色泾渭分明。与合并前相比，领导的到达度几乎没有增加——很明显，新组织的优势没有发挥出来。

为了科学地提高领导和各成员的到达度，我们在这个新体制中采取了对策。具体来说，我们根据该社交图谱显示的测量数据，分别找出了能够提高各成员到达度的最有效的对话候选人，然后让成员与其对话。

对话者的选定方法如下。从测量数据中，我们可以得知谁和谁关联以及到达度（=2 步可以到达的人数）是多少。并且还可以预测在创造新的关联时，到达度会提高多少。虽然知道自己经常和谁对话，但是仅凭个人的认识能力无法得知自己的对话者与谁关联，因此如果没有测量数据的话，就不知道自己会因新对话者的增加而增加多少个关联。通过该测量数据，我们可以超越人类的认知极限，得知通过和谁对话，可以提高多少到达度，然后利用这些信息，分别锁定每个人的对话候选人。

接着，我们按照工作相关度由高到低的顺序，为这几位候选人进行了排序。如果候选人的工作与自己的相关，那么与他关联起来就可以提高自身的到达度，从而很有可能使此人成为最有效的对话者和信息交换者。工作相关度可以通过社交图谱上的距离转化成定量数值。例如，与自己直接关联的人称为“步数为 1 的人”，与步数为 1 的人关联的人（朋友的朋友）称为“步数为 2 的人”。如果将到达某个人需要走的步数转化成定量数值，就可以把与对方在关系上的距离转化成数字。我们推测，该距离过大的话，双方在工作上的相关度就会减弱。为这几个候选人排序时，依据的就是这一距离。

做好上述准备工作之后，我们集合所有成员召开了研讨会。在研讨会上，每个小组坐在一桌（提前分好组，每组 4～5 人），针对组织

融合的目的——缩短产品开发周期，就需要采取什么措施以及自己可以做什么进行讨论。对话分为 3 个环节，每个环节约 20 分钟，每进入一个环节都会重新分组并展开讨论[在此利用的是研讨会的一种方法——World Cafe（世界咖啡）[2]]。

分组时，我们利用上述数据，对成员的搭配组合进行了精心设计。分配第一个环节的小组时，我们在能够提高到达度的对话候选人中，选择了业务比较接近的人组成一组。因为业务接近的人之间对话时容易创造话题，营造轻松的交流环境。然后，慢慢地让他们和业务不同的人进行对话，拓展视野，超越自己。之所以能这样设计，得益于可穿戴式传感器的测量结果。

该组织合并 3 个月以来，召开了 4 次这样的研讨会。我们由此成功提高了组织成员和领导的到达度，进而大幅提升了从自己周围的能力和信息中受益的可能性（即运气）。

我们将领导可以用几步到达所有成员作为该组织整合的进度指标，以前领导到达所有成员需要 5.9 步，3 个月后缩短至 3.7 步。领导的到达度提高到原来的 2 倍左右，表示现场凝聚力的三角形个数，即凝聚度也提高了 50%。

结果表明，在短短 3 个月的时间里，两个组织实现了深入融合（卷首插图 5）。这在社交图谱上也一目了然。用数字和图表来实现组织整合进度的可视化，堪称划时代的变化。

还有很重要的一点是，合并之前经常出现的推迟开发的现象不再发生。这是每位成员运气改善的结果。开发产品时，无法提前应对的技术困难和意料之外的问题是在所难免的。能否跨越这些问题，是成功的关键所在。成员的运气改善了，也就意味着在发生问题时，尽早察觉并快速解决的概率提高了。

当然，不是发生什么问题都可以保证解决。不过，概率提高带来的效果不容小觑。即使是棒球的第一击球手，打击率也只有 35% 左

右，一般的击球手是 27% 左右，但这 8% 的差距就是决定胜负的关键。商业亦然。

以成员为对象的问卷调查的结果也表明了这一点。有 9 成的人回答“更容易找别人商量事情了”“更愿意发表意见了”。有疑问立刻请教别人，是“运气”好的人的共同特征。结果就是，有 8 成的人回答“解决问题的速度提高了”。

如果产品的开发时间比计划延迟了，企业就会丧失商机，失去顾客的信任，进而导致竞争力降低。并且，如果职员为处理已经推迟的事情长时间地工作，还会陷入疲劳状态。消除开发的推迟可以打破这种恶性循环，具有重大意义。

根据到达度的指标科学地改善运气，可以为企业经营带来革命性变化。

4.9 要想抓住运气，对话质量也很重要

前面我们将通过与人相遇以改善运气的想法转化成了定量数值。但是，结识了新的朋友后，还需要找到两人对话中的连接点，以引出具体的信息和能力。其中，对话的质量至关重要。

我们再回到前面提到的史蒂夫·乔布斯的例子。因为这个实例中浓缩了可以抓住运气的本质对话。

午餐会上，乔布斯旁边坐的碰巧是诺贝尔奖获得者伯格教授，他听教授谈论了有关基因重组的内容，在此新的关联产生了。毋庸置疑，伯格教授的背后肯定与很多人和信息相关联。因此，乔布斯通过该对话大幅提升了到达度。

在这里我们注意到，乔布斯与伯格教授的专业领域截然不同，计算机与基因重组看起来也毫无关系，尽管如此，乔布斯仍然积极地听

取了伯格教授有关基因重组的谈话。据说，一旦有机会获取知识，乔布斯就会很积极地听别人说话。他有一个习惯，即使是乍看之下和自己的工作毫不相关的话题，他也会带着对说话者及其工作的好奇心认真倾听。

但仅仅这样，只不过是拓展了自己的知识面，而乔布斯并没有停留在倾听谈话、获取知识的层面。他进一步提出了一个朴素的问题——用计算机来模拟会怎样，从而将表面看来关联较少的基因研究和自己所在的计算机领域结合在了一起。起先，乔布斯主要是作为倾听者，而通过这个问题，他找到了与自己领域的连接点。就是从这里开始，才有了后来的一系列发展。

这就是对话质量的重要性，不是仅靠"关联"一词就可以阐述清楚的。接下来，我们将借助可穿戴式传感器的测量数据，对其进行深入分析。

4.10　对话即为身体活动的投接球练习

我们在第 2 章中介绍过，人在想积极解决问题并努力创新时，对话中超越基准值的快速身体运动的次数就会增多。也就是说，对话中的"活跃度"会提高。具体来说，对方说话时，我们点头附和、适时发问、抒发己见等，都会导致快速身体运动的增多。高质量的意见交换必然伴随着身体运动的增多。说到对话质量，我们的关注点往往在对话内容和对方所说的话上，但实际上，对话时的身体活动才体现了对话的质量。而对话质量可以借助可穿戴式传感器测量的身体运动来评价。下面，我们思考一下身体运动和对话的关系。

对话经常被比作投接球。一个人先发话，另一个人接话之后再发话。从这个角度来看，对话有时会表现为"语言的投接球"。

但实际上，“语言的投接球”这个表达方式并没有正确地把握对话。我们观察一下现实中的对话，就会马上明白原因。假设你是领导，要告诉下属（佐藤先生）工作调动一事。

> 我们公司正在扩大印度方面的业务，为了大幅增强该业务的运营体制，我们决定 10 月 1 日派佐藤先生常驻印度，希望您在日趋扩大的印度市场上大显身手。

实际上，在你说这句话的 20 秒的时间里，下属佐藤先生一言未发，即不存在语言的投接球。但是，他浑身上下应该都表现出了自己对这次调令的态度。可能你也在一边说话，一边观察他的反应，试图读取他用身体发出的信号。

下属发出的信号体现在眼神、表情、头的朝向、手的动作、整个身体的姿势以及微妙的动作上。

或许佐藤先生一直怀揣海外常驻的梦想，而且今后将迎来亚洲的时代，他想体验在经济不断发展的亚洲工作。因此当他听到调令后，顿时目光炯炯。或者正好相反，最近佐藤先生的妻子查出患有重病，他必须照顾妻子，因此对于此次调令正深感苦恼。

仔细想来，在这种情况下，文字信息是无关紧要的。假如佐藤先生回答:

> 要派到印度的什么地方呢?

单看文字的话，怎样理解都可以，根本看不出佐藤先生对调令的态度是积极还是不满，是犹豫不决还是惴惴不安。

但是，如果你在场的话，看一下佐藤先生用身体和语气发出的信号，就知道他的态度如何了。在日常生活中我们会通过除了语言以外

的信息来推测说话人的态度，从出生那一刻起就一直在做这种读取信号的训练了。

我们的大脑认为，对方会用“No”“不”等语言来表达消极的反应。

但实际上，在现实的场景中，你不是凭借语言来识别对方的反应的。当对方做出消极反应时，他会故意不用身体活动回应你的身体活动，以表达拒绝。而且他还会通过移开视线或转过脸去，来有意识地表达这一点。这些交流手段连婴儿都知道。你时刻都在感受对方的这些反应，并据此调整说话方式和内容。

通过对交流的分析可知，语言要素对交流的影响不足 10%，剩下 90% 多的影响来源于身体运动等非语言要素[①]。

我们一直在使用可穿戴式传感器科学地测量并研究交流中的身体运动所产生的作用。我们首先利用可穿戴式传感器中的红外线，检测出双方面对面且距离较近，然后在此基础上解析传感器中嵌入的 3 个轴向的加速度传感器信号，并测量其身体的运动情况。加速度传感器每秒能捕捉 50 次（即 20 毫秒 1 次）详细的人体运动波形，连极其微小的动作都能捕捉在内。如果能测量x、y、z这 3 个方向的波形，就可以再现此人的身体活动。

我们可以有效利用传感器测出的身体运动数据，将对话看作“身体活动的投接球”，而不是“语言的投接球”。因为当我们想向对方传达某种信息时，会下意识地使用超出基准值的高频动作。发出动作信息的人称作投手，接收动作信息的人称作捕手。很多情况下，投手也是发出语言信息的人，捕手也是接收语言信息的人。对话时，当有人发出动作和语言信息，有的人会接收，有的人则无视，还有的人会做出回应，各种各样的反应是同时发生的。这些都表现在身体运动中，而身体运动可以用传感器来测量。

4.11　有关单向交流和双向交流的研究

仅凭这些测量数据，真的可以科学定义理想的对话和会议，并测量其质量吗？答案是 YES。我们已经发现了几个指标，在此介绍一下其中最基本的指标，就是对话的“双向率”。

如果对话是用来传递信息的，或许仅通过单向交流即可，没必要进行双向交流。但是，德鲁克说：“交流不同于信息，两者是相反或者互补的。”（《管理：任务、责任和实践》）[3]。

所谓交流，听话者理解了才有意义，问题不在说话者。但是，说话者和听话者看世界的前提不同，每个人的经验不同、能力各异，对同一句话的理解也不尽相同。问题和对立因此产生，我们怎么做至关重要。

那么，我们应该怎样面对这样的对立和问题呢？比利时研究机构 IMEC 的弗兰基·卡托尔（Francky Catthoor）教授和我们进行了共同研究，教授将对待对立和问题的态度大致分为 3 种[4]。一是展开建设性②讨论，超越对立以寻求解答。二是遵从领导的意见，即“追随”（Follow）。三是对于跨越立场和意见的不同表现出消极的姿态，认为如果可以消除对立，对方就有责任改变或采取行动（自己没有责任），这种态度称作“怀疑”（Skeptical）。

在建设、追随、怀疑这 3 种态度中，通常来说建设是最理想的。卡托尔教授的交际理论表明，建设可以做到 100%，应该以建设为目标。关于追随，下属追随领导相当于组织按照指挥命令（即上情下达）行事，看起来似乎不存在任何问题，毕竟如果领导是超人，能掌握一切，所有的能力都在下属之上的话，的确也没问题。但是，现实并非如此。一般来说，下属更了解现场的实际情况，在现场的操作能力也更强。下属拥有的知识和能力需要与领导的见识和权力相配合。

怀疑，是让双方徒劳地花费时间和精力。为了说服不积极的人，需要花费大量的时间和精力。另一方面，如果无论别人说什么，你只认可自己知道和相信的事，那结果只会陷入自以为是的境地，不会发生任何变化。而且，怀疑的态度本身也不利于心理健康。

怀疑是一种削弱组织力量的重大疾病。然而，它已经成了各种各样的人的习惯。因为怀疑可以暂时给人带来轻松，并避免徒劳。从长远来看，这种习惯会腐蚀组织和人本身。

卡托尔教授的交际理论明确地将对话这种复杂现象划分为3种模式，可谓一针见血。但是很遗憾，卡托尔理论是定性的。

不过，最近这种理论有了定量依据。一桥大学的沼上教授等人在108家企业中，独立开展了大规模的问卷调查（称作“组织的重量”研究[5]）并证实：通过建设性对话解决问题，与企业的收益性息息相关（沼上教授报告中所说的“直接对决=彻底讨论，辨明是非”相当于卡托尔理论的“建设”）。研究表明，存在“重量”问题的企业和没有“重量”问题的企业相比，利润率平均高出0.2σ（σ是所调查的组织中利润率的差异值，称作“标准偏差”）。

调查同时证实，解决问题时存在追随行动和怀疑态度的企业收益性较低（在沼上教授的报告中，除了直接对决以外，还有强权、妥协、回避问题3种，而卡托尔理论中，除了建设以外，还有追随、怀疑2种，因此两者并非一一对应。但是，笔者考虑到两者的内容存在共同点，于是将直接对决和建设之外的态度作为相同内容进行了比较）。在追随或怀疑态度较多的企业中，组织的动作会变得沉重。所谓“沉重的组织”，指的是内部调整多、组织涣散（缺乏紧张感）、很少挑战新事物的组织。

对话的质量非常重要，而使用我们的可穿戴式传感器，就可以检测对话是否是建设性的。因为在建设性对话中，对话者之间的双向率会提高，而在追随、怀疑性的对话中，动作的双向率必然会降低。

4.12　根据身体运动的测定值，可以明确定义对话的质量指标

我们是这样定义“双向率”这个定量指标的。选取两个正在对话的人，每分钟调查一次两人是否存在超过基准值的快速身体运动（和测量活跃度时使用的判定基准相同），当两者都有这样的快速身体运动（1 分钟内）时，称作“双向”。相反地，当这 1 分钟内只有一个人有超过基准值的快速身体运动时，称作“单向”。有快速身体运动的一方称作“投手”，没有快速身体运动的一方称作“捕手”。我们在整个对话期间进行测量，并将“双向”占整个对话时间的比率定义为“双向率”。例如，1 个小时的对话相当于 60 个“双向”或“单向”（投手或捕手）的状态。其中，“双向”有 30 分钟的话，“双向率”即为 0.5，这样就实现了定量化。我们用这种方式得出双向率，其中，双向率高的对话是建设性的对话，反过来，建设性对话的双向率会提高。这个指标可以结合第 2 章中论述的积极解决问题与对话时活跃度的关系、活跃度的传染等，进行综合考量。

在这里很重要的一点是，虽然建设性的讨论很重要，但是现实往往不会这么理想。如果领导专制强势，你稍微表达一点不同意见，就会失去立足之地或被贬职，这时你就不得不选择追随了。又或者，你凭借自己的经验和知识，绞尽脑汁也想不出自己参与策划的项目的意义和目的，那么出席项目会议时，你就会表现出怀疑的态度。要克服这种状况并非易事。

但是，大数据和测量可能会改变这种状况。假设从大量数据中导出的“对话中的双向率和收益的关系”已在社会上广为人知，而且可以测量公司内的对话情况。那么，如果对话者可以像照镜子一样收到反馈的话，其行动应该会发生很大的变化。在前面说的领导专制强势的情况下，测出的双向率就应该极低。在收益与对话双向率的关系广为人知的前提下，如果看到自己的“对话双向率”数据极低，就不得不改变自己的行动了。

怎样才能提高对话的双向率呢？这个问题的答案，正在渐渐揭晓。我们可以直接把刚刚介绍的内容，即对话的双向率十分重要这个

认识分享给组织的成员。要让他们理解为什么追随和怀疑不可取，为什么人们动辄陷入追随和怀疑之中。

但是，仅仅这样做还不够。下面我讲一下自己的体验。在我们的职场中，可以根据对话者的不同，随时确认对话的双向率数值。根据这些数值，我发现自己和一个下属的双向率与其他人相比总是明显偏低。我在对话时已经注意了，但毫无改善。起初我以为只是因为与对方性格不同，但事实上关系不大，还存在真正的、更深层次的原因。

实际上，要想提高对话的双向率，重在设定一个有挑战性的目标，要求两人进行认真的交流。否则，不会产生推心置腹的双向讨论。

在我的例子中，我和双向率低的下属之间，没有一个有挑战性的共同目标，而只是一味地想要改善表面的对话形式，这是无济于事的。对话的质量是一面镜子，可以展示两人是否在进行挑战。在我意识到这点后，为这个下属的工作设定了一个有挑战性的目标。结果立竿见影，我和这个下属的对话“双向率”急剧上升。同时，我们也证实了定量测量是一个强有力的工具。

让成员认识到高双向率对话的重要性，可以提高对话的双向率。但是，从根本上来说，提高双向率本身不是目的，我们最好将双向率视作反映工作及人生“挑战性”的镜子、指标。也就是说，通过对话的双向率，我们可以确认相关人员是否在做有挑战性的工作。不难理解，挑战程度与企业的收益息息相关。

一直以来都是公司的人事部门采取对策，提高职员对公司的贡献，比如管理者培训、交流研修和早期选拔等。但是，这些对策究竟会给职员的微观行动带来多大的影响，我们不得而知，也没有测量的手段。至于这些对策是否和公司的宏观业绩有关，更是无人知晓。

我们可以使用新的测量技术，每天记录并分析公司所有职员间的交流数据。同时，如果将迄今为止管理的各种宏观指标（收益、接单率、顾客满意度和员工满意度等）和微观指标结合起来，就能明白为

了直接提升公司的业绩，应该按照怎样的优先度开展交流，而且还可以构建这样的交流系统。

交流要花成本。一直以来，我们都是根据个人经验和感觉决定如何分配有限而宝贵的时间、与什么人交流等。通过持续测量并分析交流与业绩的关系系统，我们可以结合当时的情况，采取对业绩最有效的交流。

从这个意义上看，对人和组织来说，可穿戴式传感器及其分析技术有望成为新时代的“镜子”。也就是说，通过这面镜子，我们可以看到自己的状态，组织也可以看到组织本身的状态。这一工具可以帮我们认真面对对话的质量、自己和公司的“运气”，而以往我们无法看到这些。这几十年来，人们崇尚合乎常理、合乎逻辑的思考方式，从而丧失了认真面对“运气”的态度，或许现在通过这面镜子，我们可以重拾这种态度。

4.13　从“运气也是实力的一种”到“运气即实力”

中国有一本古籍叫《易经》，这是有关人类命运最全面也是最早的书籍。《易经》中有一句话：

> 君子知微知彰
>
> （领导者能看出事物微小时的征兆，也能看清其显著时的特征。）

关于人类和社会，我们觉得自己看得很清楚，但实际上很多东西我们都没看到。古籍的作者说，为了看见这些隐藏的事物，我们要具备一种智慧，以看清表面看不出的变化模式和结构（关于《易经》这

一部分的作者是谁，有很多种说法。但是，曾经在 1,000 多年的历史里，人们认为这是孔子的著作。小林秀雄认为，像曾经的人们所相信的那样去回忆过去，才是面对历史应有的态度。在这里，笔者也效仿这种态度，将《易经》看作孔子的著作）。而且孔子说过，能做到这点的才是真正的领导者（君子）。这是一个普遍的真理，从 2,000 多年前至今从未改变。

在此介绍的新科学技术告诉我们，表面不易看到的人与人之间的关联和身体运动中存在着支配运气的因素。这个因素就像一面镜子，促使人们采取下一次行动。

常言道，运气也是实力的一种。回顾本章论述的内容后，你会发现“运气即实力”的说法更为准确。

不管是人生还是工作，都会被运气（概率）支配，谁也无法否认这一点。而不含运气、可以机械完成的事情，附加值一般比较低。现在，这些低附加值的工作，要么用计算机处理，要么在成本较低的新兴国家进行，从而大大节省了费用。对于在日本从事的工作而言，成败基本在于如何控制运气。

棒球的击球手时刻都在直面自己的运气。无论是多么厉害的击球手，3 次中也有 2 次失手的时候。但是，经过多次打席，实力就会清楚地反映在“打击率”这个概率数值上。在棒球的世界中，“运气才是实力”是一个常识。为了提高哪怕 1 厘的概率，棒球手们在不断磨炼自己。一郎[③]为了提高这个概率，从每天的饮食到通往棒球场的台阶上先迈哪只脚，都会有意识地加以控制。

英国作家、医生塞缪尔·斯迈尔斯在其著作《自助论》（1859 年）（*Self Help*）中写“天助自助者”，这与福泽谕吉的《劝学篇》（1872 年）一起，对明治时代的日本人产生了巨大的影响。在这里，“天”这个词表示的应该是掌握命运的关键。本章中笔者一直在论述，我们应该通过自己的努力，控制能够得到自己想要的东西的概率——这种思考方式并不新颖。

日本现在很多的大企业，都是由明治时代的人创建的，他们受到斯迈尔斯的影响，接受的是以《易经》为代表的儒学教育。

前面我们提到了被誉为“日本资本主义之父”的涩泽荣一，他在其著作《论语与算盘》中阐述了如何协调企业经营与公益的关系，同时还论述了不可避免的偶然、直面问题的态度以及通过自己的努力改善“运气”的姿态。

马场粂夫博士奠定了我所在的日立公司的技术基础，他曾指出：“学问分两种，一种是人的学问，一种是物的学问。（中略）社会万象形成于人类内心和物体运动。”（《易经新研究》1960 年）马场从小学习儒学，后来他甚至有了一个信念：《易经》论述了面对变化的原理原则，理应运用于实际的公司经营。为了管理变化，需要人和物两者合一。我们从这种想法出发，针对难以避免的产品事故，在日立建立了一种制度：当事故发生时，从人和物两方面研究并学习事物现象。我们根据米勒那幅有名的画，将这种制度命名为“拾穗者”。多年来，在拾穗者制度的支持下，日立生产出了信赖度较高的产品。我们从正面捕捉运气，并将其作为拥有 40 万人的组织行动 DNA 融入了制度之中。

非常遗憾，这几十年来，认真面对这些无形运气的态度，已经在日本企业中消失殆尽了。如今，在占卜中“运气”一词往往被理解为毫无根据的扭曲之物。甚至可以说，这导致日本的组织愈发沉重，难以盈利。

但是，我们在有关运气的科学研究中，逐渐看到了希望。将运气可视化并改善运气的新技术，正在逐渐开辟一条从根本上改变人生与经营的道路。

① 关于非语言交流，具体请参阅 Marjorie Vargas 的《非语言交流》（*Louder than Words: an Introduction to Nonverbal Communication*）。

② 原文用的是 Critical 一词，遵照词意翻译过来就是“临界性”，但日语翻译成“建设”更自然。

③ 日本棒球明星铃木一郎。——译者注

第 5 章
撼动经济的新“无形之手”

5.1　社会能否科学化

看似我们一天的行动和优先顺序可以由自己的意志和选择决定，但是研究表明，它们受到了科学规律的限制。现在我们通过获取大量的新数据，可以更加清楚地看到这一点。

就如同宇宙大爆炸和基因的发现一样，新数据为科学带来的新突破，在科学史上反反复复地发生着。看似人类行为受到了错综复杂的主观感受和情绪的支配，但实际上发挥作用的却是一直用来说明物质性质的、优美的物理定律和数理，这个发现着实令人惊讶。

本章的目的就是将该发现扩展到经济活动。试想，我们是否可以科学地认识并控制企业业绩呢?

5.2　从科学角度来看，不知“买”为何物

经济活动的基础在于我们的“购买”行为。

通过购买，面包师才可以集中精力制作面包。为了制作面包，不仅需要面粉和黄油等材料，还需要大量供自己和家人使用的生活必需品。面包师可以通过购买获得这些生活必需品。他用面包换取货币，又用货币购得面粉和生活必需品。农民在面包师不知道的地方种植小麦，其劳动成果在上述过程中，以面包的形式广泛地分配到社会上。

要想在面包师和农民之间建立合作关系，两者既不用相互认识，也不必建立信赖关系。两者的宗教信仰可以截然不同，思想也可以互不相容。不同的人的劳动成果通过购买行为相互关联，分配到社会各处。

进一步说，面包师用心制作面包，农民集中精力生产小麦，他们可以提高各自的能力、生产力和产品质量。为了更好地创造资源，并将资源广泛地分配到社会上，“购买”是一个基本要素。

让人意外的是，尽管购买行为如此重要，但实际上，从科学角度来看，人们并不知道“买”为何物。

假设今天你要和家人开车去郊外的购物中心。买完东西回家时，请看看都买了些什么，分别花了多少钱。这是由什么决定的，又是什么时候决定的呢?

仔细一想，这个问题极其复杂。希望你回顾一下自己从开始购物到结束购物的过程中接触过的人的数量和他们的工作，物品、信息的数量和种类，以及因受到这些影响所采取的种种行动。这些信息量非常庞大，我们根本无从得知其中哪一项影响了购买行为，哪一项没有影响购买行为。如果被问到今天为什么购买这件东西，人们可能会给出一个看似正确的解释。但是，这不过是牵强附会地为了解释而解释罢了。

5.3　如何从科学角度解读经济活动

当我们试图理解购买这一行为时，一般会认为人是在自己内心的动机、意识的驱动下产生了行动。经济学也是在这种立场上构建的。在经济学中，人们一直都认为，人的购买行为是由人内心的价值标准（称为“效用函数”）决定的。因此当有多个选项时，人们会选择效用高的行动。

最近，人们已经可以测量脑部活动了。人们开始更多地利用脑部测量数据，用大脑这一物质体以及在大脑内发生的神经元发火和脑内物质（多巴胺等）分泌等生理现象代为解释人的心理和意识。这也属于从人类内心寻求原因，与以往借助效用函数进行理解有相似之处。

但是，在借助可穿戴式传感器测量并研究身体运动的过程中，我渐渐开始质疑这些研究方式。因为我意识到，以往人们一直用来理解复杂自然现象的研究方法，与在人们内心寻求原因的研究方法有着根本不同。

这一想法产生的契机来自一种类推，即“人”对应“原子”，“社会现象”对应“自然现象”。第 1 章中说过，人类行为与原子的能量分布遵循同一个公式。这种对应关系不是偶然的，两者均起源于资源在构成要素间的反复交换。其特别之处在于，原子的能量象征着行动时胳膊的活动次数。

我们回顾一下按照该对应关系研究自然现象的理论。

迄今为止，自然现象一直被看作原子和周围原子团相互作用的复合系统。例如，水蒸气在一定的温度以下会变成液体（凝结），金属在一定的温度以下电阻会变成零（超导),人们是怎么理解这些现象的呢？无论我们多么详细地调查构成水和金属的氢原子和铜原子的性质，都无法解释这些现象。构成物质的原子会影响其周围的原子，同时，周围的原子团创造的“场”（电磁场等）会影响场外的原子。我们一直都是通过原子及其周围原子的双向作用来理解这些自然现象的。

这种科学方法与借助人类内心的动机、效用函数和脑部活动来说明人类行为的方法截然不同。我不是要否定这些方法，只是觉得人类

与周围的相互作用更为重要。

如果我们将成功解读了自然现象的科学方法运用于社会现象，情况会如何呢？组成社会的每一个人会影响周围的人和物，同时，此人周围的人和物形成的“场”，又会限制或推动场外人的言行举止，人类行为就是这样产生的。

人类行为产生于“人”和“Context（文脉）”（这里指的是周围的人和物所构成的环境）的相互作用。我认为，不能只把场外的人分离出去，或者仅仅将 Context 分离出去，必须将两者视为一个“复合系统”。

近年来，这种思考方式已被用于认识人类发展和教育等，称为DST（Dynamic Systems Theory：动态系统理论[1]）。DST 的根据之一是，同一个人，处于不同的状况（Context）时，会表现出不同的能力和行动。例如，有时我们会在现实生活中看到，孩子在老师面前能解开某个问题，但到了父母面前就不会解了。即人的能力和行为根据状况的不同而变化万千。

为了定量地推进该研究方法，我们开发出了在测量人类行为的基础上测量状况的技术，在世界上属于首次。利用这项技术，我们可以完整测量并科学解析购买这样的社会经济现象。

5.4　购买行为的全貌测量系统

接下来介绍一项我们团队开发的测量社会与经济现象的技术。

这项技术可以完整地测量实际店铺中发生的购买行为的全过程。我们在某家居建材商店的协助下，测量并收集了店铺中顾客和员工的大量活动数据。在我们对结果进行解析并据此调整了卖场之后，店铺营业额开始大幅提升。

用于测量的便是可穿戴式姓名牌传感器。

该姓名牌传感器和名片一样大小，可以用绳子挂在脖子上（图 2-1），重量只有 33g。传感器的佩戴者在约 2～3m 以内见面时，会通过红外线将自己的 ID 数据发送给对方，同时接收对方的 ID 数据。这时，时间信息（Timestamp）以及两人相互见面的事实都会记录在嵌入传感器内的存储器里。我们可以掌握传感器佩戴者何时、与谁见了面（参考第 2～4 章）。在本次实验中，不仅是员工和店长，顾客也配合我们佩戴了该传感器。这样一来，我们就可以记录顾客和员工的交流、员工间的对话和店长对员工的指导等。

经过设置，我们还明确了佩戴传感器的顾客和员工在店内的位置。

汽车导航系统和手机中安装的 GPS（Global Positioning System）都是有名的位置信息获取技术。但是，GPS 需要与卫星通信，在室内不能使用。

而本技术在店铺等室内也可以获取详细的位置信息。具体来说，我们提前在货架上每隔 2～3m 的位置设置一个能发送位置信息的红外线发射器（Beacon）。这一发射器可以通过红外线向周围发送位置识别号码（ID 号码）。这种红外线信号的传播距离有 2～3m 左右，所以佩戴了姓名牌传感器的顾客和员工一旦进入距离发射器 2～3m 的范围，姓名牌传感器就会接收位置的 ID 号码。将该 ID 号码与时间信息一起保存到存储器中，顾客和员工所处的地点和时间就被记录下来了。与电波不同，如果被障碍物挡住，红外线就传递不过去。因此，即使是在 2～3m 以内，也不会误测到货架后面的人。在家具建材商店的实验中，我们在大约 1,000 坪[①]的店铺中设置了 500 个红外线发射器来获取位置信息。借助该技术，我们可以收集动线信息，即顾客和员工在店内是如何移动的。甚至还可以获取很微小的信息，比如在通道中，顾客会看向两侧相对的货架的哪一边（卷首插图 6）。

该姓名牌传感器还可以通过内部嵌入的加速度传感器测出顾客和员工的身体活动。如前面所述，加速度信息在分析人类行为方面发挥

了举足轻重的作用。从传感器的摇摆模式中，可以检测出用户的步行情况，此外，还可以实现表示此人积极性的“活跃度”和对话交流类型的“双向率”（实时记录投接球：对话时自己发给对方和对方发给自己的信号）的定量化（参考第 2～4 章）。加速度传感器可以每隔 20 毫秒（1 秒 50 次）记录一次x、y、z三个轴的加速度数据。我们可以通过这些详细的活动数据，来分析姓名牌传感器的微弱摇摆和姓名牌的角度。

综上所述，借助该传感器我们可以收集顾客在哪儿、停留了多长时间、在何时何地与哪个店员进行了对话、对话属于什么类型的交流（不记录对话内容）等所有信息。

而且，我们还可以收集每个员工何时在哪个卖场、是否在库房收货、哪个员工和哪个员工进行了什么类型的交流等信息。

此外，从管理的角度来看，还要收集店长和副店长何时和谁进行了沟通、何时何地进行了工作等信息。顾客和店员行动时的动作和对话活跃度等也要收集。

实际进行实验时，我们在入口处随机选取顾客，并请他们戴上姓名牌传感器。我们向他们说明了在店内进行实况调查的目的，拜托他们在店铺购物的过程中佩戴姓名牌传感器。

结合收银台的购买记录（POS 数据）、排班和店铺布局图（哪个地方在卖什么等信息），我们可以全面收集各种有关店内购买行为和业务运营的数据，而以往这些数据是无从得知的。

很重要的一点是，通过这项技术，我们可以测出顾客从进店到出店的所有行动及其与周围环境（商品、货架、通道和员工）的相互作用。换言之，我们可以定量测量有关人类购买行为的人与周围环境相复合的经济现象。

5.5　计算机 VS 人类，通过提高销售一决胜负

对于认识购买现象而言，获取大量数据本身就是一个很大的进步。但是我们不能仅仅满足于收集数据。仅凭数量庞大、种类繁多的数据，是无法看清购买现象的全貌的。为了读取海量数据中潜藏的意义，发现提高业绩的关键，我们同时也致力于开发新的大数据专用计算机。该计算机（人工智能软件）称为“Hitachi Online Learning Machine for Elastic Society”，下文简称“H”。下面介绍一下具体的实验成果[2]。

我们首先在此商店花了 10 天时间运用该测量系统获取预备数据，并用 H 分析了这些测量数据。然后我们让人类专家和 H 竞赛，看看 1 个月后这家店铺的销售分别可以提高到何种程度。目标就是提高来店顾客在店内的消费金额（其平均值称为“客单价”）。

我们请在流通行业小有成就的两位专家组成一队，共同制定提高销售的对策。此外，我们采访了有多年工作经验的公司骨干，咨询了店长和负责店铺改善的人员，同时参考了预先录入的数据。经过这一系列准备，我们确定了水管配件和 LED 灯具等应该主推的商品群，并在店内设置广告，改善货架陈列等。

而另一方面，人工智能 H 把输入计算机中的大量数据先分解成一个个小要素，然后对这些要素进行各种搭配组合并重新合成，进而自动生成可能会对提高业绩产生影响的、庞大的数据因素群。具体来说，计算机会自动生成 6,000 个影响业绩的因素，并对这些因素与业绩统计的相关性进行全面检查。这时，我们并没有利用流通行业的常识和假说，而只是单纯地使用了数据。

经过上述检查工作，人工智能 H 提出了一个影响客单价的、出人意料的业绩因素——在店内某个特定地点站有店员。我们称之为“高灵敏度地点”。H 用定量数值告诉我们，店员在高灵敏度地点的停留时间每增加 10 秒，顾客的购买金额会平均提高 145 日元。因此，实验时我们拜托店员尽量在高灵敏度地点上停留。

1 个月后，我们再次测量这家店铺，并收集了数据。结果极富戏剧性。人类专家实施的对策对店铺业绩和顾客行为几乎没有产生任何

影响。借助我们的测量技术是可以获取顾客和店员的详细行动数据的，因此我们可以从定量角度证明对策并未奏效。

那么，人工智能 H 的成果又如何呢？我们拜托店员尽量多在 H 指出的高灵敏度地点上停留，因而店员的停留时间增至 1.7 倍。结果整个店铺的客单价竟然提高了 15%，这可以说是巨大的业绩成果。客单价提高 15% 的话，销售也会提高 15%，那利润情况怎么样呢？在增加的这部分销售中，必须扣除对应商品的采购成本。扣除采购成本后，营业利润率的增长高达 5 个百分点。日本的流通行业中，营业利润率最高是 5% 左右。营业利润率增长 5 个百分点的话，也就相当于利润增长了一倍。

有趣的是，我们需要解释店员在高灵敏度地点的停留和客单价的提升之间的关联性，可要用语言说清楚并非易事。店员在高灵敏度地点停留时，顾客在店内的走动路线会发生变化，在以往很少经过的高价商品的货架前停留的时间会增加——我们有证据可以证明这一点。但让人费解的是，既然是为了改变顾客的走动路线，那为什么要将高灵敏度地点定在远离问题商品货架的地方（实际上离得很远）？此外，后面我们也会提到，店员在高灵敏度地点停留后，店员和顾客身体运动的“活跃度”也随之提高了，这点就更难解释了。对于这些提高销售额的要素，我们即便通过实验证实了，也无法直观地说明其原因所在。因此，人类想要提前确立假说，无异于天方夜谭。对于这种人类绝对无法确立的假说，人工智能 H 却可以做到。

5.6 学习型机器大显神威的时代

H 从大量数据中导出了提升业绩的对策，下面我们就介绍一下这项划时代的技术。

过去，人类基本都是将自己想处理的功能作为计算机程序输入计算机，然后从中输出数据。由于计算机程序表示的是功能和动作的模型，因此以往这种信息处理方式可以称为“从输入的模型中生成数据”。

以往的信息处理主要是用计算机代替人工作业，以提高效率。如果将人类的工作转化成模型，那么一直以来人们花费时间和成本所做的工作，都可以用计算机自动输出了。例如，以前人们需要花几天时间来计算公司员工的薪酬并统计全国店铺的销售业绩，现在就可以让计算机来做。

一般来说，信息处理分为演绎和归纳两种。可以说，迄今为止的信息处理都属于演绎。我们重温一下演绎的定义：从一般性、普遍性的前提，得出个别性、特殊性结论的推论方法（维基百科）。迄今为止的信息处理，都是人类通过制作计算机程序，写入一般性、普遍性的前提，然后得出数据这种个别性、特殊性的结论。对于那些前提和一般规律已经明确的问题，演绎法可以发挥作用，但是对于那些前提尚且不明的问题，演绎法就无计可施了。目前，计算机在演绎处理方面，展现出了非凡的能力，而在归纳处理方面，却显得无能为力。

目前，想要有效运用大数据，需要“归纳”的能力，而这种能力正是计算机一直以来的短板。所谓“归纳”，指的是从个别性、特殊性事例中找出一般性、普遍性的规律（维基百科）。现在，在输入意义不明的大量数据后，我们需要可以明确这些数据背后的模式和规律的能力。简而言之，就是要创造一种“学习型机器”。

在学习型机器中，输入的是数据，输出的是从数据中分析得出的规律。也就是说，我们需要从输入的数据中倒推出生成数据的源头模型。如果能做到这一点，那么在前面店铺的例子中，我们就可以从大量的数据中找出最能有效提高店铺业绩的原理。

微软公司的创始人比尔·盖茨在 2004 年说过：“学习型机器的诞生，相当于 10 个微软公司的价值。”（日本经济新闻，2014 年 3 月 4 日）另一方面，从这句话中可以看出，虽然他认识到了这项技术的巨

大影响力，但是在十多年前的 2004 年，距离学习型机器的诞生，还有很长的路要走。

现如今，我们的学习型人工智能 H 将畅想变成了现实。

这种兼具演绎和归纳两种处理方式、时刻从大量数据中学习的新型计算机，将在解决复杂问题、判断状况、判断经营情况方面大显神威。

5.7　人类的假说验证分析不能用于大数据

多年来，人们一直在研究用计算机分析大量数据的技术（称为 Analytics）。这项技术和我们现在介绍的“学习型机器”有何不同呢?

一直以来，分析者都是使用擅长演绎的计算机来分析数据（Analytics）。能进行这种分析的专家被称为“数据科学家”，作为现在最受关注的职业之一，数据科学家广受世人期待。

但是，这里有一个很大的问题。分析数据本来是一种“归纳性”工作，却不得不使用“演绎专用”的计算机。为了弥补其中的差距，在分析数据时，人类必须制定一个恰当的假说。那么，人类能制定出恰当的假说吗?

我们来看一下实验店铺的实际情况吧。数据的数量庞大、种类繁多，包括顾客、店员、货架、商品、时间和行动等等。数据属性的选项过多，人们根本不知道怎么制定假说。庞大的数据中究竟包含着怎样的现象和规律，人类无从想象。

实际上，人类根本就制定不了假说。明知制定不了却非要制定，那制定出的只能是相关人员容易想到的和已知的假说。就像这次竞赛中的专家一样，只能根据对相关人员的采访、以往的经验和直觉来制定假说，也只能用数据来验证该假说。

而且，这种由分析者提出假说并验证的方式，要花费巨大的时间和精力。制定假说时，还需要咨询相关人员并调查现场。从经验来看，包括上述咨询调查在内，对分析用数据的整理工作占整个分析工作的 90% 以上。即使可以使用计算机，9 成以上的工作仍是连续的人工作业和反复的实验摸索。

这很接近工匠从事的手工业。看到以往的大数据分析现场，我们会产生一种错觉，仿佛又回到了家庭手工作坊。“分析家”“数据科学家”看似是最先进的高科技职业，但实际上他们完全身处手工业世界，靠的是师傅与徒弟的直觉和经验。那些重要的、需要人力的工作，既没有实现工业化，也没有实现计算机化。

即使花费这么多的人力，按照提前制定的假说进行分析，很多情况下得出的也都是“理所当然”的结果。这一性价比未免太低。

我们在这里做的工作，正是历史上的科学家一直在做的工作。所谓科学家的工作，指的是找出观测数据背后的规律。回顾科学的历史，这是牛顿、路德维希·玻尔兹曼、爱因斯坦、薛定谔等少数天才所做的工作。而且，这样的科学发现屈指可数。只要不改变类似于以往手工业的方法，那么即使获取了大数据，也不会有太大改变。

有了人工智能 H 这种学习型机器，人类将不再需要“Analytics”。

5.8　学习型机器会提高人类“从过去学习的能力”

要使大数据在有科学依据的情况下发挥作用，必须用计算机代替牛顿等天才的工作。当然，模仿或代替天才的灵感并非易事。

回想一下，人类曾经梦想能像鸟一样在空中飞翔，而这个梦想是通过飞机这种完全不像鸟的东西实现的。在讨论人工智能时，有的人态度强硬，认为人类智能必须通过人来重现；也有的人想法灵活，觉

得虽然起初人类梦想是能像鸟一样在空中飞翔，结果却是发明了飞机，这说明只要有方法智能地解决问题就可以了。我的想法属于后者。

有了学习型机器，计算机便可以从大量的数据中进行学习。其学习的数量和速度，远远超过人类。飞机可以用远超鸟类的速度在各大陆之间航行，而学习型机器则可以从大量数据中发现天才都发现不了的规律。

同样的事情也发生在将棋的世界。在日本一场名为“电王战”的计算机和棋手的对战中，计算机软件拥有了远超一流棋手的能力（那些勇于挑战计算机的人们，尤其是敢于挑战这场吉凶未卜的对战的专业棋手让我深受触动。我发自内心地为他们鼓掌喝彩）。

很重要的一点是，计算机能从过去的大量棋谱中学习并迅速提升能力。电王战看似是计算机和人类的对战，但其实可以说是系统性的学习方式与以往的学习方式——即从人类过去的所有智慧中学习，与从自己的体验中学习之间的对战，结果是前者胜出了。

有意思的是，专业棋手最近开始学习计算机的下法了，说不定人类会因此掌握新的能力。

如前面所述，对于 H 发现的提高业绩的对策，我们很难单纯用以往的店铺经营常识来理解。但是，经历几次之后，人类就会开始学习。因此，应该说 H 是促进人类积极学习的机器。

难道我们就不能掌握计算机那种分析并学习大量数据的能力，准确且科学地判断业务和经营吗？从这种想法出发，我们开发出了人工智能 H。“H（Hitachi Online Learning Machine for Elastic Society）”这个名字来自英国侦探小说家柯南·道尔笔下的夏洛克·福尔摩斯（Sherlock Holmes）名字的首字母。在《血字的研究》中，夏洛克·福尔摩斯初次与华生见面，就根据他的细微特征猜到他有参军经历。根据证据和状况，福尔摩斯可以倒推出犯人和犯罪情况。同样，H 也是从数据中倒推出产生数据的处理模式和模型。夏洛克·福尔摩斯说道：

> 我完全不会先入为主，而是如实还原事实真相。
>
> 最重要的是，能从诸多事实中，分辨出何为重大事项，何为附带事项。否则只会浪费精力和注意力，无法聚精会神。（《赖盖特之谜》）

这也是我们的人工智能 H 要达到的目标。

为了达成这个目标，人工智能 H 具备了以往计算机没有的特点，即可以让输入的各种数据相互结合，自动输出可能会影响业绩的大量因素。并且，在多种因素搭配组合后，会形成复合因素，其数量就更为庞大了。

一般来说，人们想要提升的目标变量数据，是店铺 1 天的营业额（日销售额）等宏观信息，而用来说明这些信息的大量数据，是表示每时每刻、每个顾客、每个店员、每个地点的微观信息。表面看来这些数据与业绩的直接关系不大，而“H”的特点便是可以填补这两者之间的鸿沟。我们将弥补微观和宏观之间差距的独立技术（已申请专利）称为“跳跃学习”（Leap Learning）。

下面我们介绍一下跳跃学习的优点。假设我们要探讨能有效提高店铺业绩的因素。我们很容易预测到，店铺的业绩取决于店铺面积（面积大的店铺营业额高）。对此也很容易分析，因为店铺的业绩与用来说明该业绩的店铺面积的“粒度”相同（数据一一对应）。如果我们将两者列个表，用电子表格进行回归分析，就可以预测出店铺面积每扩大 1 坪，营业额会增加多少。由于业绩和店铺面积是粒度相同的宏观量（每家店铺都有一个数据），因此很好分析。

在这次竞争中，H 发现的对策就属于影响业绩的微观因素，要想找到微观因素，并不像上面预测宏观因素那样简单。针对什么顾客、在店铺的哪里、什么时段、哪个员工接待顾客，才能最有效地提升业绩呢？要想锁定影响业绩的因素，就要对这些条件进行搭配组合，而

组合的数量是十分庞大的。H 具备一种引擎，可以有效地从数据中找到影响业绩的各种因素，这是以往的数据科学家分析不出来的。

运用大数据时，最大课题是如何填补宏观业绩与微观数据之间的粒度鸿沟（为了说明一个宏观数据，要动用各种微观数据的组合）。但是，据我所知，在以往的分析技术（称为统计学、多变量解析和机器学习）中，没有一项可以弥补微观和宏观的差距，输出其背后潜藏的模型。

10 年前，我们刚开始研究大数据时，就遇到了如何弥补微观和宏观之间的差距这个难题。我们研究了各种运用大数据的工作，发现这些工作都涉及微观和宏观两个方面。因此我们思考，这个难题是不是已经被攻克了？于是，我们在统计学、多变量解析和机器学习中探索，看有没有解决这个难题的技术。

然而，寻而未果。可能是因为在大数据出现之前，这些学术领域的问题就已经固定了。例如，多变量解析广泛应用于心理学的问卷调查和临床医学的药物疗效分析，我们只要研究“人”这类粒度统一的数据即可，不必关心宏观和微观的差距。机器学习应用于图像识别（图像中的脸部识别等），我们只要研究“图像”这类粒度统一的数据即可。不管是多变量解析还是机器学习，都不必处理大数据中粒度不同的微观和宏观数据。

这 10 年来，伴随大数据的出现，处理粒度不同的微观和宏观数据的需求应运而生，但是人们从来没有正面解决过这个新问题。因此，我们决定从零开始，开发大数据分析技术，即学习型机器 H。

5.9　3 种人工智能

从机器会学习这个意义上来说，人们研究开发出的人工智能全都在以某种形式进行学习。不过，根据目标方向的不同，人工智能可以

分为 3 种，分别为“运转判断型”“问答型”“模式识别型”。

其中，无需对大数据进行分析的学习型机器 H 属于“运转判断型”人工智能。具体内容如本书的介绍，比如提高店铺和呼叫中心的生产力，降低铁路和水处理设备等社会基础设施的运转成本，协助判断顾客需求和市场定位等。最近发展神速的将棋和围棋软件也属于运转判断型的人工智能。

“运转判断型”人工智能的原理特点在于，将现实世界理解为一种微观要素——集合（=集团）。这是 19 世纪下半叶奥地利物理学家路德维希·玻尔兹曼所采取的方法的延伸（章末注 1）。

该领域最近发展迅速，很大程度上得益于过去 100 年间先人构建的统计力学的体系和技术方法。而且，随着数据的大量收集，统计力学的作用也越来越得到发挥。

笔者一直关注“运转判断型”人工智能的巨大影响，致力于开发人工智能 H。

下面简单说一下其他两种人工智能，供大家参考。

很多人印象中的“人工智能”最接近第 2 种——“问答型”。问答型人工智能指的是用语言回答人的提问，当今问答型人工智能的代表就是 Google 公司的网页检索系统。最近，IBM 公司开发了一款叫作“Watson”的问答型人工智能。Watson 参加了某猜谜节目，战胜了猜谜大王，顿时成为人们议论的对象。问答型人工智能不会输出直接的判断结果，而是会提供与提问相关的信息和知识。至于使用这些信息和知识进行具体的判断，就是人类的工作了。第 3 种是“模式识别型”的人工智能，可以用计算机来识别图像、声音等数据。比如锁定照片中的人、识别人说的话等。这个领域也随着数据量的增加、计算机的高速化以及机器学习技术的高度化而飞速发展。例如，手机中安装的声音识别软件和数码相机中安装的人脸识别软件就属于该领域。

迄今为止，在人工智能和机器学习的领域中，关于“问答型”和“模式识别型”的研究很多，“运转判断型”的研究相对较少。现在大数据备

受关注，想必今后“运转判断型”人工智能也会飞速发展。

5.10　通过大数据获取利益的 3 项原则

我的研究团队从 10 年前就一直在研究如何从大数据中发现价值，那时还没有“大数据”这个词。我们的这个构想领先了世界七八年。

看到这里的读者，可能会觉得这项研究很顺利地得出了结果，但实际上，在该研究过程中，我们遇到了接二连三的困难。不过，我们的研究并非无用功。应该说，该研究有作为先行研究的意义。在困难中吸取教训，改变做法，并为此开发技术，研究适用的方法——经过这一系列努力，我们在此介绍的成果才开始显现。

在这 10 年的研究中，我们总结出了运用大数据时的 3 项原则。实际上，我们因为做了违背这些原则的事情而大吃苦头，后来我们遵循了这些原则，研究就步入了正轨。

“通过大数据获取利益的 3 项原则”如下所示[3]。

> 第 1 项原则　明确应该提高的业绩（Outcome）
>
> 第 2 项原则　广泛收集与提高业绩相关的人财物数据
>
> 第 3 项原则　不依靠假说，让计算机从数据中倒推出提高业绩的对策

在前面提到的商店的事例中，一开始我们就有意识地遵循了这 3 项原则。我们先介绍一下最重要却一直做不到的第 3 项原则。

简单来说，第 3 项原则就是让计算机制定假说。之所以做不到这点，是因为对于数据分析，人们普遍认为是由人类制定假说，然后用

计算机和数据来验证该假说。

制定假说并验证，是解决问题的正确步骤。但是，当问题中存在大数据，就不应该由人制定假说。只有由计算机制定假说，才能体现出大数据的价值。我们必须舍弃人类制定假说这个固有观念。

前面也说过好多次，人类根本无法理解大数据的全貌。别说全貌，就连其概要人类都无法掌握——这就是大数据的特征。在这种情况下，人类必定会无视大数据的价值，依靠经验和直觉制定假说。对于存在各种大数据的问题，由计算机制定假说的时代到来了。

应该由人来制定假说，这一想法给人施加了很大的压力。如果有人被分配到有大数据的项目中，并试图做些什么，就会相应地产生一些费用。而想要使用大数据的当事人，通常没有审批费用的权限。于是，当事人必须向有审批权限的公司领导进行说明。然而，很多情况下，没有假说根本通过不了审批。因此当事人会希望将上述 3 项原则普及给公司的高层，推进正确的探讨。

当然这无可厚非，毕竟以往还没有第 3 项原则需要的学习型机器。不过现在，学习型机器 H 的出现改变了这个状况。

第 1 项原则和第 2 项原则论述的是应用第 3 项原则的前提。第 1 项原则是明确应该提高的业绩（Outcome）。对企业来说，业绩尤为重要，直接关系到财务收益。

也许你认为第 1 项原则是理所当然会做到的，但实际上做不到的情况居多。更准确的说法是，一直以来，我们因为做不到这一点而大吃苦头。在大数据的事例中，人们往往会认为，既然有这么多的数据，能不能想办法用上它们呢？很多时候，人们以这样的想法为契机开始利用大数据，而这种想法单纯地作为一个契机也的确是可以的。

但是，如果我们无法推定应该提高的业绩，这种想法就会失败。因为不能推定，就无法制定工作目标。这个道理看似谁都明白，却谁都做不到，着实出人意料。之所以会这样，是因为我们在开始观察数据，并且实现了部分数据的可视化后，就会感觉很新鲜。这意想不到

的趣味性也使得顾客兴趣盎然。因此，仅仅是实现了数据的可视化，就会让顾客误以为数据是有价值的。但是，冷静地想一想，不与财务收益挂钩的东西，是没有最终价值的——这就是第 1 项原则想要告诉我们的问题。

假设我们遵守了第 1 项原则，但要遵守第 2 项原则更是难上加难。很多情况下我们会想，仅靠自己轻松得到的数据是不是就能做些什么，而不会想要广泛地收集与业绩相关的数据。

具体来说，这里有两大关卡。首先，我们往往认为，使用别人的数据时，必须要有明确的理由。而且，考虑到既然用了别人的数据，对方就会期待相应的成果，我们往往会将索要的数据限定在可以用理论说明的最低限度。这就是阻碍我们遵守第 2 项原则的关卡。这样一想，必然会认为要想得到别人的数据，必须先设定假说（结果就偏离了第 3 项原则）。但是，我们明明不知道数据中潜藏着什么，却不懂装懂地说出自己的假说和得出成果的概率，以此来索取数据——这种做法并不可行。我们希望能够正大光明地公开说出："关于假说，就让计算机根据数据来制定吧。"

再者，人、财、物等信息全都和业绩相关，很多时候，在信息系统积累的数据中，关于物和财的数据有很多，而关于人的数据却严重匮乏。如果无视这个事实并贸然推进，结局只能是无功而返。

与人相关的数据之所以重要，是因为顾客和员工的行为会大大影响业绩。

营利活动由 4 层构造组成。首先，第 1 层是应该提高的财务层，财务直接反映业绩。第 2 层是需求层。所谓需求，指的是顾客需求和购买行为，因为付款的主体是顾客，所以需求自然会大大影响第 1 层的财务。第 3 层是业务层。回应顾客需求的是业务，业务的成功与否自然会影响需求。第 4 层是设备和投资层。对设备和人才的中长期投资决定了业务的生产力、规模和质量。基础设施的准备和人才培养等就属于这一层。

人、财、物要素贯穿于这 4 层构造，尤其是在第 2 层的需求和第 3 层的业务中，人的行为有着决定性的影响。因此，我们必须针对需求和业务中有关人的信息进行深入的分析。然而，很多情况下我们都没有进行这种分析。

现在已经有很多方法可以获取有关人类的数据。人类行为数据的测量和解析是一项有偿服务（日立高新技术集团在提供这项服务，服务名称为“人类大数据云”），虽然多少要花些费用，但是必要的数据还是要获取的。在数据不足的情况下就开始分析会浪费宝贵的人力和时间，因此获取到人类数据之后再分析反而更划算。利用这些方式收集有关人、财、物的数据至关重要。

5.11　学习型机器可用于解决所有社会问题

上述 3 项原则和学习型机器可以广泛用于解决各种问题。例如，我们将呼叫中心的业务作为 Outcome，构建一个体系以广泛收集业务记录和负责人的行动数据，并输入学习型机器。这个事例我们在第 2 章就介绍过了。

又比如，我们将学校和教育机构的学生成绩、教学成果作为 Outcome，输入考试成绩的变化、学生的升学地和课堂内外师生行动的数据。这样一来，就可以在计算机上构建一个包括学生、老师和课程在内的虚拟学校模型，建立起在有限资源下将学校的教学成果最大化的行动体系。

我们再把眼光放远一点，将振兴城市经济和缓解交通拥堵作为 Outcome，广泛收集整个城市的行动数据。如果能得到居民的协助，就可以有效利用手机上的位置信息、加速度传感器中的信息，以及经济统计数据和交通信息。将这些信息输入学习型机器，就可以在计算

机上自动构建一个虚拟城市模型。基于这个模型，我们可以利用有限的预算和资源，根据数据找到促进城市发展的方法。

我们进一步把眼光放到最远，考虑如何解决整个地球的问题。例如，我们将解决地球环境问题、全球化经济发展、解决国际纠纷作为Outcome。为此，我们要收集地球上的各种数据，并将其输入学习型机器。想要输入所有数据，就需要一个超大规模的计算机和存储器。假设真有这样超大规模的计算机和存储器，就可以在计算机上构建整个地球的模型了。或许我们可以在计算机上构建有关复杂地球问题的模型，并得到一些解决思路。

5.12 人类和工作将与机器共同进化

学习型机器的出现，将大大改变社会服务和人类在社会中的作用。

回顾历史，成为20世纪经济发展原动力的是美国管理学家弗雷德里克·泰勒（1856—1915年）的《科学管理原理》。泰勒深入研究了钢铁厂中的铁铲作业。研究过程中，泰勒将工作分解成一个个过程（或动作），在每个过程中可以发现一些徒劳的工作和短时间内能改善的工作。然后他整理了每个过程的标准状态，将其写成手册，并严格按照手册推进工作。这样一来，对于那些看似只有熟练工才能做的工作，即使是经验少的工人也可以做，并且能够保证一定的质量。

泰勒的科学管理法后来也被称为“工业工程”，于20世纪被广泛应用于各种业务。这种将业务分解成一个个过程、实现过程标准化，以避免徒劳的方法至今被广泛应用于所有业务和服务中。

20世纪下半叶，人们利用计算机彻底落实了这一手段。用计算机程序写下处理步骤，就可以按照步骤处理并输出大量数据了。起初，人们用计算机处理会计事务，后来扩展至所有企业活动，开始用计算

机掌握接单、采购、制造、库存、发货和人事等一切业务流程。在这里，人们按照泰勒的思考方式将业务分解成一个个过程，实现过程的标准化，然后用计算机记录并管理每个标准化过程的状态和动作。

计算机软件一旦成形，就无法像人类一样灵活变通，我们可以利用这一点，用计算机将标准化的业务流程彻底贯彻到整个组织中，同时还可以降低以往在管理方面花费的庞大的间接费用（人事费）。

重要的是，工作方法的革新使得新机器（计算机）的产生成为可能，而机器又可以协助工作的实施，两者相互发展，共同丰富了社会。

德鲁克指出，在泰勒之后，体力劳动的生产率平均每年提高3.5%，于 20 世纪末提高到了 50 倍。这“成为 20 世纪经济和社会发展的基础”，而且“创造了我们今天所说的发达国家经济”（《已经发生的未来》）。

我们将按照泰勒的思考方式共享最有效的方法并提高了能力的人称作“Human 2.0”。其最大特点在于，将人进行了“标准化”。与之相对，我们将以前通过分工推进“专业化”的形态称作“Human 1.0”（图 5-1）——这是亚当·斯密描述的工作者（Worker）的特点。

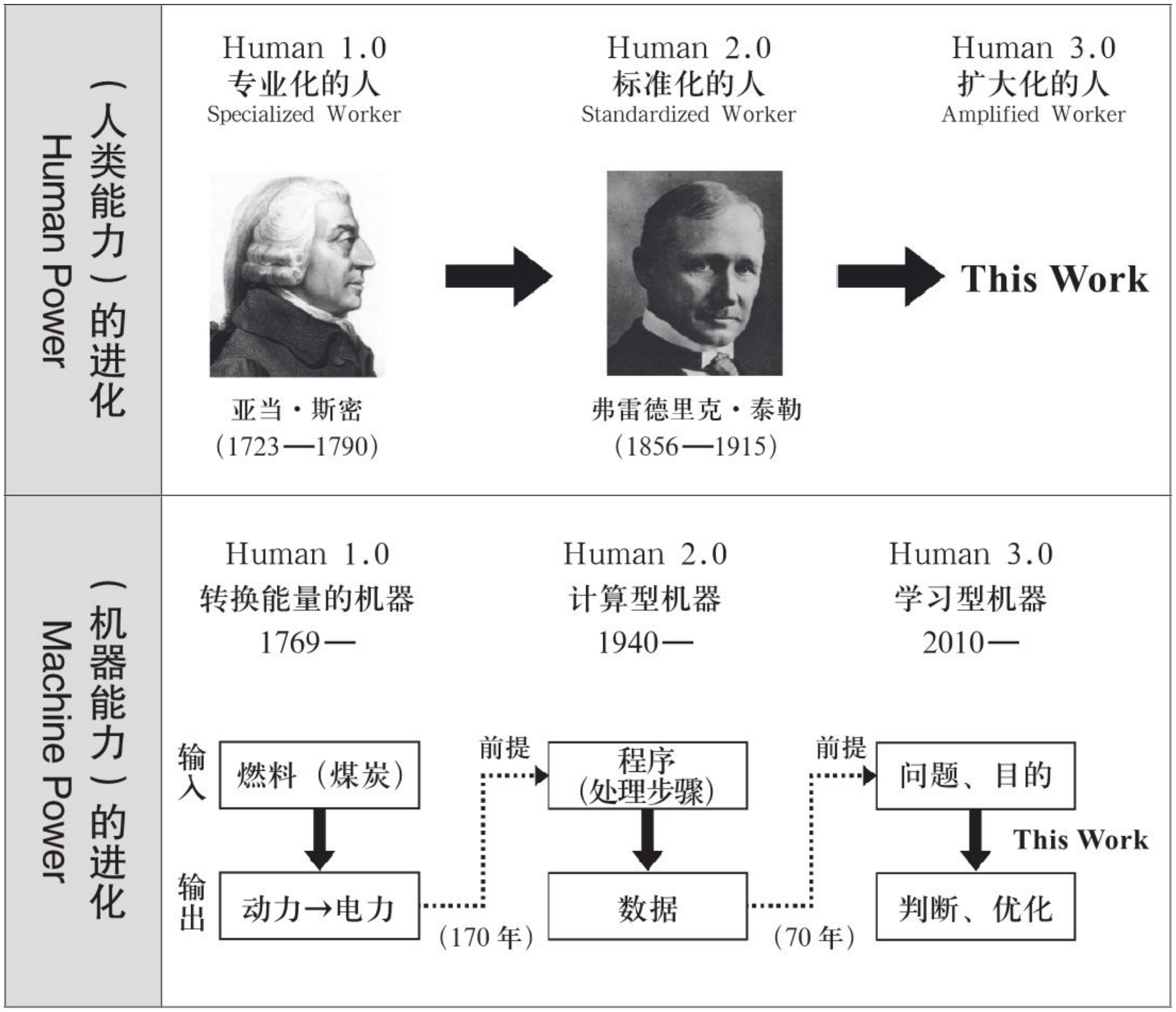


图 5-1 Human Power（人类能力）和 Machine Power（机器能力）的共同进化促进了生产力的提高。分工制造工具、深化专业的人类（第 1 代）共享并学习优秀人员的技术（第 2 代），跨越时空，从全部现实中直接而自主地学习，以提高自己的能力（第 3 代）。

一直以来，企业信息系统都在支持 Human2.0 的工作。但是，现在的投资回报率很快就要达到极限了。这就是我们在此介绍的第 3 代机器（信息系统）出现的背景。

如前面所述，如果德鲁克高度评价了泰勒的工作，那么正确指出其工作极限的也是德鲁克。很遗憾，在很多服务中，即使实现了业务流程的标准化并整理成手册，对生产力的提高也是有限的。

例如，仅靠手册无法提高护士和商场售货员的工作效率。护士除了看护患者这一本职工作外，还必须写文件、与相关人员开会、负责各种调整事务等。商场的售货员也一样，除了为顾客推荐商品、促使顾客购买的本职工作以外，还要写文件、调查库存、留意配送状况等。在复杂多变的情况下，对这些工作优先程度的判断和时间分配方法，是无法写成手册的，只能靠护士和售货员本人的判断。

德鲁克指出，从这个意义上来看，护士和售货员不是手册工作者，而是知识工作者——原本知识工作者这个词就是德鲁克创造的（经常有人狭隘地将知识工作者理解为白领。但是，这个词的创造者德鲁克将知识工作者用于最广泛的对象，列举了护士、售货员和汽车修理工等例子）。当今发达国家的很多工作都属于服务型产业，并逐渐成为知识型工作。

对知识工作者而言，重要的是设定工作目的和目标。之所以这样说，是因为只有明确了目的，才能在复杂多变的情况下灵活而准确地做出判断。如何面对变化，决定了企业的兴衰。对此，德鲁克描述道："我们知道变化是不可避免的。引起变化正是企业的主要功能之一。"（《管理的实践》）

基本来说，通过磨炼人的能力，可以提高其灵活应对企业变化的适应力。德鲁克在著作中阐述了这一点。并且，德鲁克反复论述了知识工作者本身对结果负有责任，强调了"自我约束，持续学习"的重要性。

如今出现的学习型机器以及运用该机器的信息系统有力补充并提高了人类的学习能力，或许也可以由此提高生产力。我们称之为第3代机器，即 Human 3.0。其特点在于将人类的能力"扩大化"（图5-1）。就像将ERP（整合业务软件包）这样的计算机运用于 Human 2.0的"标准化"一样，在 Human 3.0 中，计算机可以协助人们持续学习大量的数据，而这些数据是人类自己看不过来的。这样一来，仅凭人的经验就可以做出准确的判断了。而且还可以让人们做出以往业务手册中没有记录的灵活判断，即使商业状况（流通和供需等状况）改变，

也可以适应变化。这与以往那种即使情况改变，也继续同一种做法的习惯形成了鲜明的对比。

以往人们普遍认为，业务的标准化和手册的制定是提高业务的应有姿态，用户和员工也一直在迎合固定的手册和机器。虽然实现了业务流程的标准化，并借助计算机严格遵守这些标准，但实际上,在很多需要应对复杂多变的状况的服务行业中,这种做法并不可行。应该说，人们甚至担心该做法会阻碍业务的推进。

而在第 3 代机器的工作中，不是人类去迎合机器和流程，而是让机器迎合人类。在变化的环境中，机器会为自主做出判断并对结果负责的人们提供支持。

今后，计算机将发展成主动学习数据、理解本书所论述的人类身体与社会的规律和制约的机器。用这样的学习型机器武装起来的知识型工作和服务，可以时常通过数据从历史和最新情况中学习，灵活适应变化，并自主地催生变化。德鲁克梦寐以求的知识型工作的理想形式就有望实现了。

5.13　人类应做之事与不必做之事

有人可能觉得，计算机都进化到这种地步了，岂不是没有人类要做的工作了。其实不然，只有人类能做的工作还剩 3 项。

第 1 项，学习型机器不能设定问题。它只不过是在收到问题后，使用数据提供准确的信息和判断。人类要做的是明确需要解决的问题，执行学习型机器做出的判断。

第 2 项，学习型机器只适用于可以用数值来表示目的、有大量相关数据的问题。但即使我们身处未知的状况，也只能往前走。即使目

的地模糊不明，没有过去的数据，我们也只能在迷雾中摸索前行。在这样的状况下，做出决策就是人类的工作。

第 3 项，学习型机器不对结果负责。对结果负责是人类固有的能力。这时人类要做的是考虑第 1 项、第 2 项的限制，判断是否应该使用学习型机器，如果可以使用学习型机器，就需要定义应该解决的问题，并为机器提供合适的数据。不管是否使用学习型机器，对结果负责的都是人类。责任归属于人类，也就意味着工作和技术更是以人类为中心。

今后，那些学习型机器擅长的工作，想必会迅速从人转移到机器。比如对收到问题（目标可以用数值表示的、有大量相关数据的问题）的解决方法进行思考和判断的工作。代表性的例子是软件处理程序（算法）的开发工作。目前人们仍将该工作视为高度智能型工作之一，但学习型机器运用过去的大量数据自动生成算法的时代即将来临。

在将棋软件方面，以前是人把将棋棋谱作为算法编入其中。但最近，机器学习代替了人力劳动，因此将棋软件的水平实现了飞跃发展。

在声音识别和图像识别的开发工作中，也出现了完全相同的情况。多年来，人们认真研究声音韵律和音节结构，并将其编成算法，渐渐改善了声音和图像识别率——这是数不胜数的优秀研究者们一直在做的工作，而且一直被视为高度智能型工作。但是，近几年机器通过学习数据编写的识别算法[研究这种模式识别型问题的方法被称作深度学习（Deep Learning）法]很快就超越了人类编写的算法。这导致机器开始代替人类进行算法开发工作。

一直在从事这类工作的人，应该尽快将重心转移到创设问题的工作上。继续从事这种适合机器来做的工作，风险太大。

随着学习型机器的出现，人类应做之事和不必做之事发生了巨大变化。我们理应将这种变化看作人类和机器之间新型合作关系诞生的

过程。通过将合适的问题交给学习型机器，人类解决问题的能力会飞速提升。能有效利用这种方法的人或组织，与不能有效利用这种方法的人或组织将逐渐拉开差距。

5.14　新的“无形之手”将为世界带来新的“财富”

大量数据和进化后的学习型机器带来的不只是利益。人们在遵循这 3 项原则追求利润时，无形中还会为社会带来“共鸣”和“幸福”。

具体就以前面店铺购物的例子进行说明。该店铺确定了提高业绩这一目的，并输入大量数据，用主动学习数据的人工智能 H 倒推出了提高业绩的模型。根据 H 得出的结果，店铺让更多的店员站在高灵敏度地点上，最终使业绩提高了 15%。

这家公司之所以要改变员工的配置，是为了提高利润。但实际上，配置的改变不仅提高了业绩，还提高了员工和顾客的活跃度（超出基准值的快速身体运动的时间占比）。正如第 2 章中所说，活跃度提高与幸福感的提升、积极性的提高息息相关。经过详细调查后发现，业绩之所以会提高，是因为通过活跃的身体活动，员工的积极性和顾客的活跃度提高了。

人们往往认为，业绩的提高得益于改变配置带来的店内顾客流动的变化。但我们观察数据后发现，原因在于整体接待顾客时间的增加。耐人寻味的是，顾客被接待的时间长短与该顾客的购买金额并没有直接关系（从统计结果来看，没有必然的相关性）。不过，当顾客在店内看见自己周围有很多店员接待顾客的场面，该顾客的购买金额就会增多。接待顾客的直接效果在于向顾客传达其不知道的信息，间接效果在于顾客看见其他顾客在和店员积极对话，会感觉十分热闹。相比直接效果，间接效果对业绩的影响更大。实际上，从加速度传感器的测量结果可知，员工待客时的活跃度提高了。而顾客看见待客场

面后，停留时间也会随之增加，进而购买金额提高。数据表明，仅仅将店内的待客频率提高 10%，客单价就会提高 92 日元，仅仅将待客时的顾客活跃度提高 10%，客单价就会提高 68 日元。

像第 2 章中介绍的那样，人的“幸福”取决于与他人的共鸣和行动的积极性。不是说有共鸣或积极行动了就能轻易得到幸福，而是说有共鸣或积极行动本身就是人类幸福的原形。因此，当我们使用大数据来获取利益时，也会在无形之中获得与人的共鸣、积极性和幸福感。

这种现象不仅存在于店铺。正如第 2 章中介绍的那样，在呼叫中心的电话运营业务中，从数据中锁定提高销售的主要因素并进行改善，将提高员工的共鸣、积极性和幸福感。而且，以往人们认为与业绩无关的、休息时的员工活跃度提高 10% 的话，接单率就会提高 13%，这既是提高业绩的对策，也是提高员工幸福感的对策。

在 18 世纪资本主义黎明时期，英国道德哲学教授亚当·斯密用“看不见的手”一词表示自由经济的特征。也就是说，个人在追求自身经济利益的过程中，财富会在无形之中分配给社会，整个社会也随之富裕起来。

同样的事也将发生在运用大数据的第 3 代 Human 3.0 上。也就是说，运用大数据追求个人利益时，会出现一只超越“看不见的手”的、新的“数据的无形之手”。越是使用大数据追求个人利益，“数据的无形之手”就越会在无形之中引导社会走向富裕。这样一来，以往与经济价值没有直接关系的人的“共鸣”和“幸福”等，也与经济价值联系起来了。

正因为大数据和计算机纵观了多种因素之间的复杂依存关系，我们才得以发现这种联系。让人印象深刻的是，比起人类自己，计算机得出的答案甚至更关心人类。而且，计算机不会像人类那样因过度相信自己有限的经验而抱有偏见。一直以来，人们往往认为“追求经济效益”和“追求人类应有的充实感”是相互对立的，但是数据和计算机将两者联系在了一起。

亚当·斯密写过两本书，《国富论》和《道德情操论》。前者论述了经济富裕的本质，后者论述了人类应有的生活方式。这两者合二为一，形成了一种理论。斯密主张，追求经济效益和追求人性这两者应该相互配合，共同发展。然而，继斯密之后，人们普遍开始根据有限的数据单纯追求经济效益，而将追求人性抛诸脑后。斯密想说的是，经济效益和人性不是对立的，而是相关的。随着大数据和智能计算机的出现，斯密的这种想法终于成为可能。

注 1

玻尔兹曼成功地从微观气体分子运动中导出了宏观气体的性质。我们称其为统计力学，是物理学的一部分。这一知识体系从 20 世纪 80 年代开始应用于人工智能领域。如果使用该统计力学体系，即便无法预测微观构成要素的个别活动（从物理上来说就是个别粒子的运动），也可以随机获得整个群体的性质（这里可以用第 1 章中介绍的玻尔兹曼定义“熵”的概念来解释）。该体系的所有性质都会反映在基本方程式中（比如我们经常用到的将自由能表示成温度和体积的函数方程式），该方程式使判断优先程度（例如，在物理体系中，可以根据自由能的大小在多个候选状态中预测哪个状态稳定且实际可行）成为可能。

① 日本面积单位名，相当于 3305.7 平方米。——译者注

第 6 章
社会和人生的科学将带来什么

6.1　在濑户内海的直岛描绘未来

前面我们一直在介绍，通过收集和利用人与社会的定量数据，科学技术的地平线不断推进，现在扩展到了社会和人生。以往被人们狭隘地认为是理工类的科学技术，开始有了更广泛的社会意义。

伴随大数据的出现而产生的新科学今后将何去何从，又将给社会和我们的人生带来什么？利用这个新的可能性，我们应创造出怎样的新社会？

这个问题没有正确答案。准确地说，答案要靠我们今后创造。但是，答案不会不请自来，为了创造答案，我们必须集合科学、技术、产业和社会的各种智慧。为此，我们推进了一项企划：特意将平时没有交集的各个领域的专家聚集起来,一起讨论今后应该完成的重大挑战。

2010 年 3 月 6 日，32 名专家聚集在濑户内海一个名叫“直岛”的小岛上，他们几乎都素未谋面。直岛被开发成了一座艺术岛屿，安藤忠雄先生在此设计了一座清水混凝土建筑，置身建筑内外，大家可以看到奇形怪状的造型以及用沙子做成的美国国旗被蚂蚁蛀食后的样子。

这 32 人是在日本国内尖端技术、人类科学和商业最前沿的超一流关键人物。笔者组织这些评论家们进行了为期三天两夜的共同活动。

他们抵达会场后，中午吃了便当。因为大都是初次见面，所以场面有些僵。毕竟他们的领域各不相同，平常也没有交集，甚至连使用的语言也不一样。但是，通过“World Cafe”（约 4 个人一个小组，不断改变小组成员的组合，让他们进行交流）等形式，未来蓝图逐渐形成语言。这些专业领域各异，甚至没有共通的基本用语的人，开始了知识的融合。他们连日秉烛夜谈，想法超越了以往的技术和商业，扩展至未来。2 天后，我们要挑战的 21 世纪的课题转变成了重大挑战——“直岛宣言”[1]。

6.2　以社会为对象的科学迅速发展

我们先来讨论一下在直岛实施的这场研讨会的大背景。

现在正值科学技术与社会关系发生变化的时候。过去，“科学→技术→用于社会（服务化）”这种线性模型很普遍。按照这一流程，即使我们在基础研究方面有了科学发现，运用到社会上一般也需要花费 10 年以上的时间。

不只是时间的问题，现状是与科学、技术和服务社会相关的人分别处于完全不同的共同体中，互无往来，彼此没有机会共享知识。

不过，这一状况正在发生变化。我的周围开始了一种良性循环：通过运用在各种社会服务现场收集到的数据，可以科学地理解社会。而另一方面，科学理解社会又会使新的服务成为可能。本书一直在介绍的就是这点。本书内容完全偏离了线性模型，提出了一种新模型，即科学确立、技术开发与服务社会三者并驾齐驱。

为什么会出现这种新的模型？首先，收集数据并从中发现科学的环节全部实现了自动化，因此从科学技术到服务社会的进程加快了。在以往的科学中，大学研究室完成一项工作要花 1～3 年（相当于研究生取得博士学位的过程），而新模型都是以日、周为周期，两者的速度有着天壤之别。

以往只有研究者才从事的“科学”，今后将逐渐面向成千上万的人。例如，有了大数据和人工智能，店铺负责人就可以利用数据自行找到提高业绩的对策。

自从任何人都可以使用检索网站之后，获取知识的速度飞速提升，而这种情况进一步发展，就将产生上述新状况。提高速度固然重要，但更重要的是，遇到具体问题时可以利用数据，立刻解决问题。

今后，我们有望利用每日在全世界的各种现场中积累的数据，与人工智能对话的同时科学且高速地解决问题。

6.3　将服务与科学融为一体的数据呈指数增长

直岛研讨会还有另外一个背景，即数据积累速度的加快。今后，随着技术的发展，计算机会越来越小。这提升了测量世界的密度和规模，还有可能提高积累并运用数据的能力。

上述状况必然会加快技术科学和服务（包括企业服务和公共服务）的融合。今后，或许会迎来科学技术和服务平等合作以促进技术革新的时代。我们称之为“服务和科学的融合”。

通过服务和科学的融合，服务推动科学的进步，科学促进服务的发展，两者共同进化，共同发展。

服务和科学共同进化的关键在于对现实世界的大数据的积累、扩大。一直以来，我们收集了 SNS 等网络人际关系信息和网站链接的相

关信息，而今后我们将重点收集现实世界的数据。现实世界的数据不仅存储于计算机的数据库，并且每天都在增加。在本书介绍的可穿戴式传感器的例子中，现实世界的数据呈指数增长，增长速度大约是 1 年 4 倍。

数据规模之所以扩大，是因为数据收集对象的规模扩大了，而这同时又增加了新服务的价值。

数据规模的扩大会促进传感技术和数据收集技术的发展。并且，系统和运营成本的下降，会进一步促进数据收集规模的扩大，提高社会认知度。收集的数据一旦开始发挥作用，理解数据收集和积累价值的人也会增多，从而可以对数据收集进行积极反馈。

在共享数据的同时，服务进步和科技发展实现了共同进化，我们称之为“数据的指数增长定律”。迄今为止，有很多促进技术进步的定律，其中，“摩尔定律”格外有名，即集成电路上可容纳的元器件数目每隔 18 个月便会增加一倍。摩尔定律论述的是，计算机的心脏——微型芯片（集成电路）的性能和成本每年都会呈指数增长。但是，如今只看微型芯片一个硬件是片面的。今后我们应该关注的，是安置在社会各处的计算机收集和积累的社会实测数据。

从社会活动中积累的数据，在以每年 3 ~ 4 倍的速度迅速增长。“数据的指数增长定律”也许会代替摩尔定律，在近 50 年内对社会和经济的发展发挥作用。

摩尔定律使得业界（半导体行业、设备行业、材料行业、计算机行业、软件行业、金融行业）和研究所可以共享未来愿景，共同创造未来。因而近 50 年来，摩尔定律逐渐成了发展以技术为主导的社会的指导原理。

我们在这里提倡的“数据的指数增长定律”，超越了以往摩尔定律覆盖的行业，或许可以成为社会各行各业的人们共创未来的基础。

我们所收集数据的规模可以视为用来收集社会数据的集成系统的规模。也就是说，通过摩尔定律我们可以统计出容纳到 1 个芯片上的

集成装置在整个社会的使用率是多少。当晶体管变成原子大小时，摩尔定律就会走向终结（关于终结时间有多种说法，早的话是 10 年～20 年以后）。另一方面，数据的指数增长定律之后仍会继续存在。它很有可能超越芯片，成为表示技术高度化全貌的定律。

6.4 重大挑战“直岛宣言”

如果可以利用大数据实现服务与科学的融合，那么应该构建一个怎样的社会呢？要想描绘出这一蓝图，服务、科学和技术这三者必须密切配合。但现状是，这三者没有机会对话。正因为如此，我们才策划了本章开篇的直岛研讨会，将这 3 个世界的一流专家聚集在一起，加速推进上述新动向。

我们有幸请到了一流的专家们，从 2010 年 3 月 6 日到 8 日，进行了整整两天的研讨会（主题为“服务、商业科学和新技术”,日本电子信息通信学会主办）。我们利用 World Cafe 等形式，让他们加深讨论，并总结了今后这三者要配合解决的 5 类 10 个重大挑战，将其命名为“直岛宣言（Naoshima Manifesto）”。

“直岛宣言”展望了大数据和应用大数据的科学技术即将开拓的未来。本章是最后一章，我们在此介绍一下“直岛宣言”。

6.5 直岛宣言

NM1 相互感知

一直以来，在设想计算机的未来蓝图时，人们经常会想到马克·维瑟提出的 “ubiquitous （普适）”概念[2]，即“计算机将融入生活的各个角落”。但是，仅靠这一概念，计算机无法激发出人类的潜力。

今后，我们需要由计算机理解并鼓励“热情”“共鸣”等激发人类行为的主要因素。这在直岛宣言中统称为“Affective”[3]。

NM1.1 Affective·Service

首先，提高人类的热情和共鸣的服务（称作“Affective·Service”）会愈发重要。

以往，服务的价值在于用技术代替人工作业，以减轻人的负担。而有效运用大数据的 Affective·Service 已然超越了减轻负担的范畴，甚至可以支持人类发挥个人潜力。这需要我们进一步深化服务，不仅像以往那样单纯追求“便利”，还要理解并支持“生存意义”“信念”和“梦想”。因此，基于大数据的定量人类科学将成为一项重要要素。

NM1.2 Affective·Technology

实现 Affective·Service 的技术基础是 Affective·Technology。该技术是计算机针对共鸣、热情和幸福进行理解、定量、测量、记录、分析、通信和共享的技术。要想实现 Affective·Technology，必须确立从电子学到通信、信息处理等领域的各种新技术。我们想建立这样一项技术：利用大数据理解人们所处的状况，实现社会关注点和相关性的可视化与定量化。

其中，最基本的技术便是将人类的终极目的——“幸福”定量化的技术。

NM2 合力

20 世纪是世界规模的通信网络得以建立，竞争日趋激烈的时代。另一方面，很多传统价值被视为落伍，因而备受轻视。其中也包括人们相互合作、尊敬、帮助的价值。今后这些价值可能会通过大数据的应用，以新的形式复活。21 世纪的新技术反而会强化这些被遗忘的价值。

NM2.1 灵活的组织

未来是不可测的。因此，对社会来说，组织是否有能力应对难以预测的未来至关重要。而大数据有望推动组织应对未来。我们是否可以利用大数据，打造一个灵活、强大、自律的组织呢?

20 世纪初期，很多活动都是以个人和家庭为单位进行的。而 100 年后，基本上所有的社会活动都是以组织为单位进行的，即组织成了社会的基本单位。然而，多次有人指出，作为企业基础的阶级型组织容易僵化，适应环境变化的能力差等。在环境变化中，很难一直把焦点放在复杂但重要的问题上[4]。我们想构建一个 21 世纪的组织体系，让各种成员发挥各自的能力，朝着共同目标努力。

为此，我们必须重新审视那些在组织运营方面被认为是理所当然的事情。例如，我们将汇报、联络、商量的重要性作为协调组织上下的常识教给了成员。但是，有人指出，今后在此基础上，还要加上参与、联系、互助。我们致力于建立统筹个体与整体，享有共通视点的组织。大数据有可能为组织开辟一条通往科学与工学的道路。

NM2.2 组织的保健

各种各样的人跨越民族、年龄、性别、文化和能力的差异，在全球范围内相互合作——今后这点会愈发重要，但同时组织产生问题的风险也提高了。对于身体的异常，我们可以利用各种检查仪器和客观数据，同样地，对于组织存在的问题，我们能否构建一个利用数据实

现科学预防、诊断和治疗，以及成员自我管理的体系？我们能否通过这种维持组织和共同体健康状态的体系，促进各类人的持续合作？我们将这一挑战称作“组织的保健”。我们通过从大量数据中导出的科学原则和反馈，努力跨越对立，推动各类人的协力合作。

NM2.3 新终身雇佣体系

大数据还可以应用于劳动和雇佣形态方面。数据应用带来的最大可能性是，会超出千篇一律的刻板数据和管理手册的限制。我们是不是可以利用忠实反映事实的数据，采取灵活的方式应对实际情况？

放到社会中，在超越刻板规定的意义上，超越退休制这种千篇一律的管理方式有着重大的冲击力。

下面举一个其他领域的例子。对于工业用的汽轮机等设备，员工需要遵守定期更换零件的死板规定。而现在可以根据有关实际工作条件和使用环境状况的大量数据，灵活地改变更换零件的频率，因而可以有效利用资产，降低成本。

同样地，关于社会人才的更替（也可视作社会体系中的一种零件更换），我们是否也可以采取更为灵活的方式呢？归根到底，我们应该构建一种体系，让有热情、能力和意愿的人工作一辈子。也就是说，我们想找寻实现新型终身雇佣制的可能性。

一直以来，劳动都被视为赚钱的手段。因此，人们一直认为，降低工作负荷、提高报酬才是对劳动者好。但是，很多科学研究表明，工作是充实感的源泉，工作挑战才是最好的报酬。伊能忠敬和葛饰北斋的例子告诉我们，挑战无关年龄。我们致力于超越旧制度，利用数据构建符合 21 世纪“知识劳动”“服务”“革新”的雇佣和就业体系。

NM3 悉心培育

即使是在科学技术发达的现代社会，灾害、事故、恐怖袭击、纠纷、经济波动等威胁也从未间断。而大量数据及从中发现的科学，有可能会发展为科学对抗威胁的基础设施。

NM3.1 安全的社会体系

大量数据及从中发现的科学可以预测威胁，为建立安全且更经济的社会基础设施（水、农业、气象、交通等）奠定基础。

为了推进社会基础设施中各种新技术和新方法的开发，我们需要一个实验现场来制定并验证假说。为此，我们想构筑一个国家级别的社会实验平台（验证实验的场所），通过获取大量数据并用于实验，奠定有科学依据的社会体系和服务事业的基础。

NM3.2 无国界的风险管理组织

我们需要一个能广泛收集并分析科学数据，确保世界安全以及风险管理的组织。我们暂时称之为“安全支援队”。灾害、事故和气候变化带来的损害，对世界来说是很大的风险。除了防卫，日本还有其他方法为世界安全做出贡献，并提高存在感吗？通过大量数据及从中发现的科学知识，说不定可以找到主导世界安全的道路。

NM4 科学技术的重新构建

为了在社会上应用大数据，我们必须重新构建一种与之相符的科学技术方法论。

NM4.1 人类与组织的数据存储

为了促进社会和人类科学的发展，同时作为一种社会资产，构建全人类的数据存储中心至关重要。具体的应用方法尚未确定，但是各个领域的研究者和社会改革家在使用这些数据的过程中，或许会逐渐明确其应用方法。

作为先行案例，我们用可穿戴式传感器获取了大量数据，允许在非营利研究中使用这些数据，并设立了以振兴人类科学为目的的组织——“世界信号中心”（笔者兼任中心主任）。

NM4.2 大数据应用技术的标准化主导

为了让各领域的人们都能使用（感知、解析、预测）大数据，我们必须实现应用界面等的规格化、标准化。那么，日本是否可以创造出主导世界的动向呢？再加上前面说的数据存储，如果这两项得以实现，将给社会各领域带来巨大革新。

NM5 经济的重新构建

大数据的应用还会改变我们对经济的看法。或许可以在保护自然资源（树和水）和人们的幸福感的基础上，确立新的经济评价“尺度”（经济指标）。可能也会影响人们对全球问题（南北问题等）的认识和这些问题的发展趋势。这种评价指标是不是可以超越表面的经济指标，接近其背后的限制和约束条件呢？根据这一指标，科学认识世界经状况的时代有望到来。或许可以增加为各国和个人做贡献的机会，促进世界的发展。

6.6　总结——人类旺盛的生命力

前面就是直岛宣言的10个项目。虽然直岛宣言是在短时间内以研讨会的形式总结出来的，但这是由来自不同领域的32位专家集中精力探讨得出的，包含了认识未来的重要因素。通过利用大数据的科学技术，我们可以用不同于以往的形式，科学地解决重要的社会问题。

在很大程度上，科学技术的发展和进化参考了植物的生长。植物一方面在维持自身基因，另一方面又在与环境相互作用，即兴决定自己的具体结构——其出发点就在于“种子”。“学习型组织”的泰斗彼得·圣吉先生指出：“种子本身不具备树木生长所需的资源，资源存在于树木周围的生长环境中。但是，种子提供的是决定性的东西，即树木开始形成的‘场所’。种子在摄取水分和营养的同时，实现了生长过程的组织化[5]。”

为了培育21世纪，使其成长为一棵参天大树，相当于“种子”的直岛宣言和相当于能量循环过程的数据指数增长定律，或许可以为解决复杂的社会问题提供指导。

后　记

在本书结尾之际，我想介绍一下我是如何将传感器和数据用于自己的人生的。

德鲁克指出，要成为高效的知识工作者和管理者，关键在于时间管理。具体来说，德鲁克建议详细记录并分析自己使用时间的方法，以此找到有效利用时间的方法。

他表示，重点在于实时记录：

> ……不是之后根据记忆来记录，而是实时记录下每时每刻。
>
> 必须持续记录时间，每个月定期查看记录结果。（中略）
>
> 必须查看记录，重新探讨并制定每天的日程。半年以后，你肯定会发现自己被工作牵着鼻子走，不得不在一些琐事上浪费时间。
>
> 使用时间的方法可以通过练习得到改善。但是，如果不在时间管理方面不断努力的话，只能被工作牵着鼻子走。
>
> （彼得·德鲁克《卓有成效的管理者》）

我本来打算按德鲁克所说的亲身实践一下，但出乎意料的是，在现实生活中很难做到实时记录。

可穿戴式姓名牌传感器解决了这个问题，堪称最强工具。姓名牌传感器可以实时记录会面、地点、环境音量、精力集中度、身体姿势、温度和亮度等信息。我根据这些详细记录，在第二天早上记录下昨天什么时间做了什么。到了周末，就可以重新审视过去 2 周的信息，再次探讨并制定时间利用方法。

具体来说，我根据德鲁克的观点探讨了以下 3 项内容。第 1 项，找到并排除“没必要做的工作”。第 2 项，找到“其他人也可以做的工作”，并让他去做。第 3 项，通过实际记录，发现“自己可控的自由时间”惊人地少，对这部分时间进行汇总，并将其用于可以产生重要成果的事情。

这是提高我的工作效率和人生品质的原动力。

在每天查看这些数据的过程中，我开始思考，能不能自动生成符合当日状况的建议。经过几年努力，我们已经成功研制出这一建议系统，并称之为“生活信号”[1]。我每天都在使用这个系统。它就像专门的私人随身顾问，可以随时咨询，我甚至无法想象没有它的生活。

下面简单介绍一下“生活信号”的思考方式。首先，它要用可穿戴式传感器系统地捕捉人生和生活中细微变化的征兆。人生和生活中存在连本人都察觉不到的变化，这种变化会反映在睡眠和步行时间的增减数值上。为了将这些变化提取出来，我们从传感器数据中锁定了代表生活变化的 6 个特征量，分为多于前一天和少于前一天这两类，由此得出了 64 种（2^6）生活变化的模式。经历这 64 种生活模式时，回顾一下自己当时考虑欠佳的事，然后当场记录下为自己制定的建议。这样一来，下次再经历此模式时，就可自动唤醒当时为自己制定的建议。要经历所有模式，需要几年的时间。经过长期努力，我终于构建起针对所有模式的建议系统。我建立了一个系统化的建议数据库，可以应对独一无二的人生中可能发生的所有变化（有趣的是，对我的研究内容感兴趣的人用了这一建议系统后，都非常满意）。

如此一来，我们成功开发出了名为“生活信号”的系统，可以在考虑最近生活模式变化的基础上，生成针对当日的建议。

该系统的最大特点在于，可以有效利用科学技术，提高人类从过去学习的能力。实际上，我现在写的这篇文章，也有效利用了建议，即今天“不管情况有多紧张，都要稳步前进，坚定地迈步”。我本来有些犹豫该不该写这篇后记，但还是按照建议选择了坚定前行。该系统提供的建议大大改变了我的人生。

如果只是写借助该传感器进行的生活管理及其乐趣，也需要一本书的量，我之后会写。总之，现在社会和人生与科学技术的关系正在发生变化，而且已经具体影响到了我每天的人生。之所以能写出这本书，很大程度上得益于用传感器进行的时间管理。而且传感器的使用，也大大促使了我和下面提到的朋友们相识。从这个意义上来看，本书也可以说是运用大数据创作出来的。

如果读者读了本书后，可以对目前正在发生的变化产生一些新的想法，而这些想法又能成为振兴日本的契机之一，笔者将不胜荣幸。

在诸多朋友的支持和协助下，我终于完成了这本书。在此表示衷心的感谢。

在这里无法一一列出作为测量对象的众多组织成员，如果没有他们的协助，就没有这本书的诞生。在此对他们表示衷心的感谢。其中，在呼叫中心的测量和分析方面，承蒙长谷川智之先生、金坂秀雄先生的协助；在组织改革和工作空间的适用方面，承蒙谷内田孝先生、黑田英邦先生、杉本有俊先生、佐藤直基先生的协助。在此表示感谢。

在有关幸福的共同研究中,有幸得到了美国加州大学河滨分校的索尼娅·柳博米尔斯基女士和乔·钱塞勒先生的协助；在有关心流状态的共同研究中,有幸得到了克莱蒙特大学的米哈里·契克森米哈赖先生和中村先生的协助；关于姓名牌传感器及利用该传感器在美国和德国进行的验证实验,有幸得到了麻省理工学院的桑迪·彭特兰先生、乔·帕拉迪索

先生、汤姆·马龙先生、埃里克·布里涅卢福森先生、彼得·葛洛先生、石井裕先生、Sociometric Solutions 公司的本·瓦贝尔先生和丹尼尔·奥尔古因先生的协助；关于人与人之间的合作，有幸得到了 IMEC 的弗兰基·卡托尔先生的协助；关于身体行为的理解，有幸得到了东京工业大学的三宅美博先生的协助；关于计算机和信息的意义，有幸得到了东京经济大学的西垣通先生的协助。在此表示由衷的感谢。

此外，直岛宣言是在荒川文男先生、荒宏视先生、伊藤晶子女士、内山邦男先生、梅室博行先生、大石基之先生、冈田健一女士、甲斐康司先生、金田康正先生、仓田成人先生、黑田忠广先生、斋藤敦子女士、樱井贵康先生、佐藤直基先生、妹尾大先生、高桥真吾先生、高安秀树先生、高安美佐子女士、竹内健先生、野村恭彦先生、西田佳史先生、滨崎利彦先生、广濑佳生先生、藤岛实先生、前田英行先生、增田直纪先生、松冈俊匡先生、安本吉雄先生、山口裕幸先生、吉本雅彦先生、鹫田祐一先生的共同努力下完成的。除了直岛宣言以外，他们在其他方面也为本书的创作提供了诸多启发，在此表示感谢。

本书介绍的研究开发项目，还得到了日立研究开发团队、事业团队以及在此无法一一列出的人们的指导和协助。现借此机会，表示对他们的感谢。其中，森胁纪彦先生、荒宏视先生、佐藤信夫先生、渡边纯一郎先生、大久保教夫先生、早川干先生、胁坂义博先生、辻聪美先生、秋富知明先生、福间晋一先生、栗山裕之先生、田中毅先生、爱木清先生、河本健先生、山下春造先生、堀井洋一先生、藤田真理奈先生与我一起推进了本书中的研究，是论文的共同作者，在这里再次表示感谢。

我们将可穿戴式传感器技术应用于多个领域，逐渐形成了本书中记述的内容。在这方面，有幸得到了大林秀仁先生、久田真佐男先生、松坂尚先生、日置范行先生、泷勉先生、须崎喜久雄先生、柴田修达先生、小野贵司先生、石桥望先生、一关阳平先生、佐藤一彦先

生、浅田直行先生、竹内香织先生、荒木桂一先生、清水健太郎先生的帮助。

此外，多年以来，在与佐藤彰先生、后藤久雄先生、渡部武先生、过世的佐佐木伸先生的讨论中，我也获得了很多创作本书的启示。在此衷心地表示感谢。

在艰难的创作过程中，草思社的久保田创先生长期鼓励我，并协助我将本书修改得更加通俗易懂。在这里由衷地表示感谢。

最后感谢一直支持我并给予我灵感的妻子史子和女儿麻子、柚子。

参考文献

第 1 章

[1] T. Tanaka, S. Yamashita, K. Aiki, H. Kuriyama, K. Yano, Life Microscope:Continuous Daily Activity RecordingSystem with a Tiny Wireless Sensor,*2008 International Conference on Networked Sensing Systems (INSS 2008)*, pp162-165, 2008.

[2] K. Yano, The Science of HumanInteraction and Teaching, *Mind, Brain and Education*, Volume 7, Issue 1,pp19–29, March 2013.

[3] 矢野和男，渡边淳一郎，佐藤信夫，森胁纪彦.大数据的无形之手：商业和社会现象能用科学调控吗（「ビッグデータの見えざる手：ビジネスや社会現象は科学的に制御できるか」）.日立评论,95 卷 6/7 号,pp432-438,2013.

[4] 笔者参考大泽文夫《大泽流手工统计力学》（『大沢流　手づくり統計力学』名古屋大学出版会）一书中介绍的、用粒子的移动导出玻尔兹曼分布的方法，将此方法普遍化，应用于社会行为研究方面。

第 2 章

[1] K. Yano, S. Lyubomirsky & J.Chancellor, Sensing happiness: Can technology make you happy? *IEEE Spectrum*, pp26-31, Dec. 2012.

[2] S. Lyubomirsky, *The how of happiness: A new approach to getting the life you want*, New York, Penguin Press（2008）（中文译本《幸福有方法——12 大幸福行动，让你的幸福增加 40％》——译者注）

[3] S. Lyubomirsky, K. M. Sheldon,D. Schkade, Pursuing happiness: The architecture of sustainable change,*Review of General Psychology* 2005,Vol. 9, No. 2, pp111–131.

[4] Special Issue, The value of happiness: How employee well-being drives profits, *Harvard Business Review*, January-February 2012.

[5] H. J. Wilson, Wearables in the workplace, *Harvard Business Review*,September, pp23-25, 2013.

[6] J. Watanabe, M. Fujita, K. Yano, H. Kanesaka, T. Hasegawa, Resting Time Activeness Determines Team Performance in Call Centers, *ASE/IEEE Social Informatics*, Dec. 2012.

渡边淳一郎，藤田真理奈，矢野和男，金坂秀雄,长谷川智之.定量评价呼叫中心的职场活跃度给生产力带来的影响（「コールセンタにおける職場の活発度が生産性に与える影響の定量評価」）.情报处理学会论文杂志，54(4)，pp1470-1479，2013.

[7] A. Pentland, The New Science of Building Great Teams, *Harvard Business Review*, April, pp60-70, 2012.

[8] T. Akitomi, K. Ara, J. Watanabe, and K. Yano, Ferromagnetic interaction model of activity level in workplace communication, *Phys. Rev*. E 87,034801, March 2013.

[9] J. Fox, *The Myth of the Rational Market*, Harper Collins, 2009.

（中文译本《理性市场谬论：一部华尔街投资风险、收益和幻想的历史》——译者注）

第 3 章

[1] Y. Wakisaka, K. Ara, M. Hayakawa, Y. Horry, N. Moriwaki, N. Ohkubo, N. Sato, S. Tsuji, K. Yano, Beam-scan sensor node: Reliable

sensing of human interactions in organization, *Proc. 6th Int. Conf.Networked Sensing Systems*, pp. 58–61, 2009.

K. Ara, N. Kanehira, D. Olguín Olguín, B.Waber, T. Kim, A. Mohan, P. Gloor, R.Laubacher, D. Oster, A. Pentland, and K.Yano. Sensible Organizations: Changing our Business and Work Styles through Sensor Data.*Journal of Information Processing*.The Information Processing Society of Japan.Vol. 16. April,2008.

[2] A. L. Barabasi, *Nature* 435, 207-211, 2005.

A.L. Barabasi, Bursts, Dutton, 2010.

（中文译本《爆发：大数据时代预见未来的新思维》——译者注）

[3] T. Nakamura, K. Kiyono, K.Yoshiuchi, R. Nakahara, Z. R. Struzik, Y.Yamamoto, Universal scaling law in human behavior organization, *Phys.Rev. Lett*., 99, 138103, 2007.

[4] 三宅美博《关于医疗和护理服务中的环境创造与共同创新的企划调查》（「医療·介護サービスにおける場づくりと共創的イノベーションに関する企画調査」）．研究开发计划“问题解决型服务科学研究开发计划”平成 22 年（2010 年）批准项目企划调查完成报告书

野泽孝之，三宅美博．共创环境的评价.测量和控制，vol.51，no.11，pp.1064-1067（2012）

[5] 矢野和男，渡边淳一郎，佐藤信夫，森胁纪彦.大数据的无形之手：商业和社会现象能用科学调控吗（「ビッグデータの見えざる手：ビジネスや社会現象は科学的に制御できるか」）.日立评论，95 卷 6/7 号，pp432-438，2013.

[6] M. Csikszentmihalyi, *Flow: The psychology of optimal experience*, Harper& Row, New York,1990．（中文译本《当下的幸福：我们并非不快乐》——译者注）

[7] K. Ara, N. Sato, S. Tsuji, Y.Wakisaka, N. Ohkubo, Y. Horry, N.Moriwaki, K. Yano, M. Hayakawa,Predicting flow state in daily work through continuous sensing of motion rhythm, *INSS'09: Proceedings of the 6th International Conference on Networked Sensing Systems*, pp145-150, 2009.

第 4 章

[1] Lynn Wu, Benjamin N. Waber,Sinan Aral, Erik Brynjolfsson, and Alex(Sandy) Pentland, Mining Face-to-Face Interaction Networks using Sociometric Badges: Predicting Productivity in an IT Configuration Task, *Proceedings of the International Conference on Information Systems*. Paris, France.December 14-17 2008.

[2] 朱安妮塔·布朗(Juanita Brown)，戴维·伊萨克(David Isaacs).《世界咖啡：咖啡对话创造未来》（『ワールド·カフェ：カフェ的会話が未来を創る』）．香取一昭，川口大辅译，Human Value.

（中文译本《世界咖啡：创造集体智慧的汇谈方法》——编者注）

[3] P. F. Drucker, *Management:Tasks, responsibilities, practices*, Harper Busines, New York,1973.

（中文译本《管理：任务、责任和实践》——编者注）

[4] 由比利时 IMEC 的弗兰基·卡托尔（Francky Catthoor）教授提供的信息。

[5] 沼上干，轻部大，加藤俊彦，田中一弘，岛本实．《组织之"重"：日本企业组织的再次检查》（『組織の<重さ>：日本的企業組織の再点検』）．日本经济新闻出版社.

第 5 章

[1] K. W. Fischer, T. R. Bidell,(2006). Dynamic development of action and thought. In W. Damon & R. M.Lerner (Eds.), *Handbook of child psychology* (6th ed.,pp. 313–399).Hoboken, NJ: Wiley.

[2] 森胁纪彦，大久保教夫，福间晋一，矢野和男．利用人类行为的大数据发现提高店铺业绩的关键（「人間行動ビッグデータを活用した店舗業績向上要因の発見」）．日本统计学会杂志 SeriesJ 43(1)，69-83，2013．

[3] 矢野和男，渡边淳一郎，佐藤信夫，森胁纪彦. 大数据的无形之手：商业和社会现象能用科学调控吗（「ビッグデータの見えざる手：ビジネスや社会現象は科学的に制御できるか」）.日立评论，95 卷 6/7 号，pp432-438，2013.

第 6 章

[1] 矢野和男，广濑佳生，竹内健，野村恭彦．描绘 21 世纪的科学技术与重大挑战（「21 世紀の科学技術とグランドチャレンジを描く」）．日经 Electronics，pp65-75，2010 年 8 月．

[2] M. Weiser, The Computer for the 21st Century. *Scientific American* , September 1991 .

[3] 梅室博行．《情感与品质：提供感情经验的商品与服务》（『アフェクティブ·クォリティ：感情経験を提供する商品·サービス』）．日本规格协会．

[4] G. Hamel, Moon Shots for Management, *Harvard Business Review*, pp91-96, February, 2009 .

[5] P. 西克（P. センゲ）．《即将出现的未来》（『出現する未来』）．高藤裕子译，野中郁次郎监译，讲谈社．

后记

[1] 矢野和男．生活记录经验:传感器改变人生（「ライフログ経験:センサが人生を変える」）．信息处理，50(7)，pp624-632，2009.7.

出版后记

时间，幸福，运气，财富，这些我们总也无法掌控的东西，如今通过大数据分析及人工智能就可被操控。

本书中，作者利用新的传感器技术和人工智能研究了超过 100 万天的人类行为大数据，从中发现了提高个人能力、企业生产力和社会发展的隐秘规律。从胳膊的运动规律发现人类活动的极限，从而更有效地分配时间；通过测量幸福度，推动自己采取积极的行动提高幸福感；根据“到达度”与合适的人接触，用有质量的对话增加遇到好运的机会；让机器学习替人类解决各领域的问题，走向更加充实又富裕的新人类社会……

通过本书你会发现，人类行为与数据的关系从未如此紧密，不论是什么性格、职业，你的行为都遵循着统一的“U 分布”。正如书中所说，人类行为看似受到了错综复杂的情感的支配，实际上发挥作用的却是优美的物理定律和数理法则。在人类行为方程式中，我们能够为自己的行为找到更理性的依据，从而更加智慧地提高自我能力，推动社会变革。

服务热线：133-6631-2326　188-1142-1266

读者信箱：reader@hinabook.com

2018 年 7 月

Table of Contents

原因与结果的经济学

前　言

第1章　如何不受无稽之谈的蒙蔽 因果推理最根本的思考法

何谓“因果关系”和“相关关系”？

判断因果关系的三个要点

证明因果关系需要“反事实”

没有时光机就制造不出反事实吗？

用“最贴切的值”替换反事实

只有“可比较”的组才能替换

想象不出正确的反事实就会被无稽之谈蒙蔽吗？

COLUMN 1 巧克力消耗量越大，诺贝尔奖获奖人数越多？

第2章　定期接受代谢综合征体检就能长寿吗？ 因果推理的理想形态——随机对照试验

“实验”能证明因果关系

随机分组的必要性

“代谢综合征体检”与“长寿”是因果关系吗？

何谓“在统计学上具有显著性”？

定期接受体检并不能带来长寿

投入1 200亿日元税金的代谢综合征体检

“医疗费用自付比例”和“健康”是因果关系吗？

兰德健康保险实验的结果

除贫困阶层以外，提高自付比例对健康状况没有影响

COLUMN 2 整合多项研究的“元分析”

第3章　男医生比女医生更优秀吗？ 利用与实验类似的偶发现象进行自然实验
用现有数据重现与实验类似的环境
“医生性别”与“患者死亡率”是因果关系吗？
女医生负责的患者死亡率更低
“出生体重”与“健康”是因果关系吗？
出生体重较重的婴儿更健康
COLUMN 3 被动吸烟会增加心脏病的患病风险吗？
第4章　最低工资与就业之间存在因果关系吗？ 排除趋势影响的双重差分法
模仿实验的“准实验”
前后比较毫无意义
不能使用实验前后测分析的两个原因
去年的销售额为“反事实”时，前后比较才有效
实验前后测设计的改良版——双重差分法
双重差分法成立的两个前提条件
“保育所数量”与“母亲就业”是因果关系吗？
增设认可保育所不会提升母亲的就业率
“最低工资”与“就业”是因果关系吗？
提高最低工资不会减少就业
COLUMN 4 “不快点睡觉，妖怪就要来了”是正确的教育方法吗？
第5章　看电视会导致孩子学习能力下降吗？ 利用第三变量的工具变量法
广告费打折，该怎样利用
工具变量法成立的两个前提条件

“看电视”与“学习能力”是因果关系吗？
看电视可以提高偏差值
“母亲的学历”与“孩子的健康”是因果关系吗？
母亲上过大学，孩子更健康
COLUMN 5 增加女性管理层成员能促进企业成长吗？
第6章　和学霸做朋友，学习能力会提高吗？ 关注跳跃的断点回归设计
利用“49人店铺”与“50人店铺”的差异
断点回归设计成立的前提条件
“朋友的学力”与“自己的学力”是因果关系吗？
和学霸在一起，也无法提高自己的学习能力
“医疗费用自付比例”与“死亡率”是因果关系吗？
提高老年人医疗费用的自付比例对死亡率没有影响
COLUMN 6 “激素替代疗法”的陷阱
第7章　上好大学收入就会更高吗？ 组合相似个体的匹配法
找出相似的店铺
整合多个协变量的“倾向得分匹配法”
倾向得分匹配法成立的前提条件
“大学偏差值”与“收入”是因果关系吗？
上偏差值高的大学，收入也不会增长
COLUMN 7 商务版随机对照试验“A/B测试”
第8章　便于分析现有数据的回归分析

假如现有数据都不适合用来评估因果关系……
画一条表示数据的“回归线”
用“多元回归分析”排除混杂因素的影响
COLUMN 8 因果推理的发展史
补论1　了解分析的有效性和局限性
内部效度与外部效度
随机对照试验也有局限性
补论2　因果推理的五个步骤
后　记
参考文献
简单统计学
序言
第1章 模式、模式、模式
混杂效应
选择性报告与谎报
易受欺骗的本性
无论文，不生存
统计显著性膜拜
第2章 不再神奇的超级畅销书
去最好的学校
投票人数越多越好？
一醉方休
放下遥控器
请原谅我的直白
只有弱者留下来
被损坏的飞机
畅销书的秘密
第3章 被误传的谋杀之都
马萨诸塞州谋杀之都

请在我家后院开一座采石场
第4章 新的经济学上帝
政府债务临界点
相关性等同于因果关系？
堕胎会减少犯罪吗？
是我的错
第5章 扬基队的门票真的划算吗？
我的天哪
虚惊一场
让他们吃蛋糕吧
有弹性的数轴
将恶作剧翻倍
学生谎言
扬基队的门票很划算
图像的制作艺术
第6章 美国有多少非裔职业运动员？
蒙提·霍尔问题
一个名叫佛罗里达的女孩儿
条件概率的混淆
假阳性问题
罕见病问题
达特茅斯三文鱼研究
第7章 辛普森悖论
某人的悖论
我要再来一杯咖啡
第8章 状态火热的雷·阿伦
小数定律
一项篮球研究
小沃尔特·雷·威廉姆斯
投掷马蹄铁

保龄球
第9章 胜者的诅咒
均值回归
西克里斯特的愚蠢
古老的谬误何曾消亡？
道指落榜生
冠军窒息
寻找学院院长和灵魂伴侣
第10章 如何转变运气？
第11章 德克萨斯神枪手
癌症聚集恐慌
第12章 终极拖延
死亡凹陷和尖峰
经过进一步研究
午夜的月饼
漫长的告别
第13章 黑色星期一
火箭科学
巴斯克维尔猎犬
五行噩运
花押决定论
第14章 点球成金
棒球迷信
如果你给我一个D，我就会死（die）
糟糕的出生月份
名人堂的死亡之吻
第15章 特异功能真的存在吗？
超感知觉
霍迪尼的挑战
一个家庭对超自然现象的着迷

第16章 彩票是一种智商税
目之所及
我们都将为IBM工作
股票是一种不错的投资
跑赢大盘（或者说为什么我喜爱包装工队）
傻瓜四股
反向头肩
如何（不）中彩票
不会破裂的泡沫
南海泡沫
伯克希尔泡沫
真实股价
第17章 超级投资者
宽客
收敛交易
我只看数据
在推土机面前捡硬币
闪电崩盘
第18章 增长的极限
穷途末路
第19章 何时相信、何时怀疑
被模式诱惑
具有误导性的数据
变形的图像
缺乏思考的计算
寻找混杂因素
手气好
均值回归
平均定律

德克萨斯神枪手
当心经过修剪的数据
缺乏理论的数据仅仅是数据而已
缺乏数据的理论仅仅是理论而已
美好的出生日
出版后记
如何用数据解决实际问题
序章 我在日产学到了什么
让戈恩社长点头称赞的管理工具
给出“答案”的期限是3～4个月
如何在实际工作中使用数据
第1章 解决问题，你需要“流程”
第一个重点是“划定范围”
从锁定原因到研究对策
开始之前的准备工作
数据分析需要“假设”
用框架来查缺补漏
开端决定了结论的质量
开始以后很难再扩大视野
思考问题之“外”的问题
「解决问题的故事1」
问题：汽车销量下滑，怎么办？
第一步：明确计划与现实之间的差距
第2章 分解数据，找到“问题的关键”
“趋势”视点和“快照”视点
“WHAT型假设”
表示大小的“平均值”
“好用”背后的陷阱

平均值不能代表总体
了解“中位数”
“油炸豆腐”和“天妇罗面渣”，哪个更好吃
平均值所掩盖的真相
乌冬面和荞麦面的教训
用“波动”的视点给平均值做补充
标准差为什么“用不上”
“变异系数”的魅力
视觉也是“感觉”
「解决问题的故事2」
第二步：“分解”数据，锁定问题的关键
第3章 采用交叉视点，锁定“原因”
表示二者关系的“相关系数”
Excel的小妙招
锁定原因也需要“假设”
相关系数的四大优势
找到相关分析的着眼点
奖金被用到哪里了
用Excel加载项进行批量分析
利用矩阵排列优先顺序
不要随便编故事
「解决问题的故事3」
第三步：建立WHY型假设，关注影响客户忠诚度的要素
第4章 制定对策，要依据“方程式”
10秒钟完成一元回归分析
关注相关系数的平方
注意事项及应用事例

看相关系数还是看斜率
「解决问题的故事4」
第四步：通过一元回归分析，发现车型B和车型C的不同问题
第5章 用数据讲故事
解决问题的故事
把解决问题的过程展现出来
在组织中使用数据的价值与难点
你能用数字推翻众人的理解吗
更上一层楼（高级技能简介）
后 记
出版后记
人生新算法：用人工智能解读时间、幸运与财富
前 言
第 1 章 时间能否自由使用
1.1 人类行为有规律性吗
1.2 能否根据主观意志自由利用时间
1.3 支配万物的能量守恒定律同样适用于人类
1.4 通过“Life Tapestry”可以俯瞰人生
1.5 从胳膊的活动次数中发现惊人规律
1.6 递降分布统治社会之谜
1.7 反复移动就会出现 U 分布
1.8 我们在各个时间点之间调配“胳膊运动”
1.9 即使不知道微观状况也能预测宏观状况
1.10 时间的利用方法受到规律限制
1.11 “经常动的人”=“有工作能力的人”吗
1.12 把握各频带的活动预算，充分利用所有频带

1.13 没有干劲是因为活动预算用光了吗
1.14 熵是什么？是表示杂乱的量吗？
1.15 自由的牢狱——正因为自由人类才遵守规律
1.16 人类活动的极限可以用热力学公式表示
1.17 人的自由与限制
第 2 章 测量幸福
2.1 能否用科技控制人类的幸福
2.2 幸福的心理学——“积极心理学”
2.3 提升员工的幸福感有利于提高公司收益
2.4 传感器可以测出幸福感
2.5 解读行动中隐藏的符号
2.6 休息时活跃的对话有助于提高生产力
2.7 身体运动会传染，幸福也会
2.8 身体运动活跃的职场的优点
2.9 打造活力职场是一项重要经营项目
2.10 我们也要考虑 IT 会降低生产力
2.11 通过幸福科技创造幸福指标
第 3 章 求“人类行为的方程式”
3.1 人类行为中存在方程式吗
3.2 方程式究竟是什么
3.3 与人的再次见面遵循普遍定律
3.4 以见面概率为基准思考，则时间的流逝各不相同
3.5 1/T定律也适用于回邮件等其他行为
3.6 行动持续越久越停不下来
3.7 1/T定律与 U 分布相同
3.8 记述人类行为的方程式
3.9 将主观概念转化为客观数值

3.10 测量最优体验=心流
第 4 章 认真面对运气
4.1 偶然是不可控的吗
4.2 运气源于与人的相遇
4.3 将运气和相遇转化成理论和模型
4.4 到达度真的是衡量运气好坏的指标吗
4.5 运气好的人在组织中处于什么位置
4.6 领导的指挥能力与现场的自律并不矛盾
4.7 通过数值化，从语言的束缚中解放出来
4.8 通过控制“到达度”，成功实现组织整合，防止开发延迟
4.9 要想抓住运气，对话质量也很重要
4.10 对话即为身体活动的投接球练习
4.11 有关单向交流和双向交流的研究
4.12 根据身体运动的测定值，可以明确定义对话的质量指标
4.13 从“运气也是实力的一种”到“运气即实力”
第 5 章 撼动经济的新“无形之手”
5.1 社会能否科学化
5.2 从科学角度来看，不知“买”为何物
5.3 如何从科学角度解读经济活动
5.4 购买行为的全貌测量系统
5.5 计算机 VS 人类，通过提高销售一决胜负
5.6 学习型机器大显神威的时代
5.7 人类的假说验证分析不能用于大数据
5.8 学习型机器会提高人类“从过去学习的能力”
5.9 3 种人工智能

5.10 通过大数据获取利益的 3 项原则
5.11 学习型机器可用于解决所有社会问题
5.12 人类和工作将与机器共同进化
5.13 人类应做之事与不必做之事
5.14 新的“无形之手”将为世界带来新的“财富”
第 6 章 社会和人生的科学将带来什么
6.1 在濑户内海的直岛描绘未来
6.2 以社会为对象的科学迅速发展
6.3 将服务与科学融为一体的数据呈指数增长
6.4 重大挑战“直岛宣言”
6.5 直岛宣言
6.6 总结——人类旺盛的生命力
后 记
参考文献
出版后记